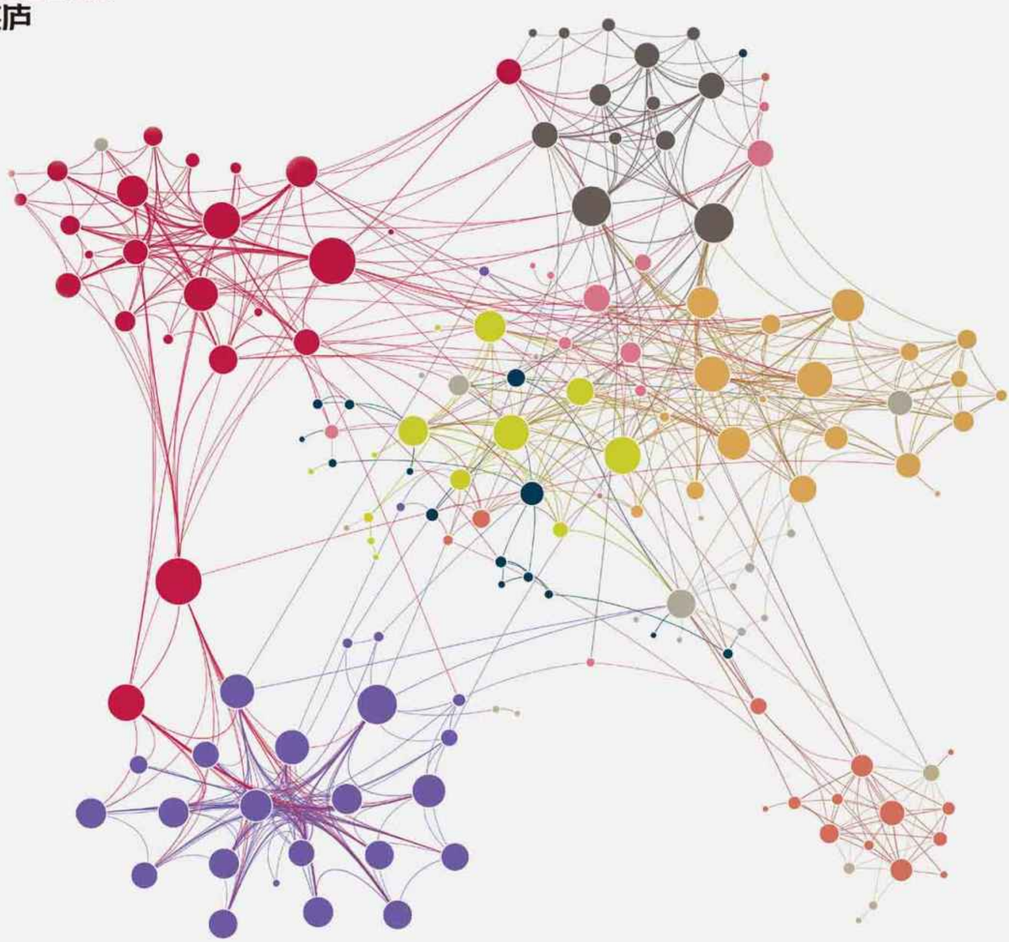
CHEERS
湛庐



复杂网络权威
首次系统揭秘
网络科学的
重磅之作

巴拉巴西
网络科学
NETWORK SCIENCE

[美]
艾伯特-拉斯洛·巴拉巴西 著
Albert-László Barabási

沈华伟 黄俊铭 译

河南科学技术出版社

VISIT...

LANZAROTE Caliente.COM

版权信息

本书纸版由河南科学技术出版社于2020年2月出版

作者授权湛庐文化（Cheers Publishing）作中国大陆（地区）电子版发行（限简体中文）

版权所有·侵权必究

书名：巴拉巴西网络科学

著者：艾伯特-拉斯洛·巴拉巴西

电子书定价：152.99元

Network Science by Albert-László Barabási

Copyright © Albert-László Barabási

All rights reserved.



ALBERT-LÁSZLÓ BARABÁSI

艾伯特 - 拉斯洛 · 巴拉巴西

全球复杂网络研究权威

“无标度网络”创立者

从被拒稿的天才物理学家到诺贝尔奖热门人选



1967年3月30日，艾伯特 - 拉斯洛·巴拉巴西出生于罗马尼亚哈尔吉塔县（Harghita）的一个匈牙利族家庭。他的父亲拉斯洛·巴拉巴西（László Barabási）是历史学家、作家，也是塞克勒博物馆的馆长。因为这一得天独厚的便利，童年时代的巴拉巴西得以自由徜徉在博物馆的图书室和各类收藏品中，甚至能够查阅只有极少数历史学家才能够查阅的文件。

年少时，巴拉巴西一度梦想成为雕刻家，但在高一时，因为在当地物理奥林匹克竞赛中获胜，所以整个高中阶段他都将兴致转向了科学与数学领域。自然而然地，当1986年考大学时，他进入位于罗马尼亚首都的布加勒斯特大学（University of Bucharest）攻读物理与工程学专业。在此期间，巴拉巴西开始进行混沌理论（chaos theory）方面的研究，并成功发表了3篇论文。1989年，巴拉巴西和他父亲移民匈牙利。两年后，他开始在匈牙利首都布达佩斯的罗兰大学（Eötvös Loránd University）攻读硕士学位，而且在还没毕业的时候，就被列入了美国波士顿大学的物理学博士招收计划。获得博士学位后，巴拉巴西在IBM托马斯·沃森研究中心（Thomas J. Watson Research Center）参加了为期一年的博士后研究。而这个阶段，也是巴拉巴西第一次对网络产生了巨大的研究兴趣，并将今后的研究方向确定为网

络科学。

虽然在学术之路上一直顺风顺水，但是巴拉巴西在1995年撰写的第一篇关于网络的论文，却连连被几家核心期刊拒稿，因为对方认为：“这和我们有什么关系呢？”但是，他并没有气馁。1999年，巴拉巴西大胆地提出了一个惊人的网络模型——无标度网络模型，该模型后来被命名为“BA模型”（Barabási-Albert model），并一举颠覆了网络的本质，成为众多后来者研究的理论基础。

今时今日，现在的人们已经深刻地认识到，复杂网络不只和我们有关，而且关系极为密切。经过多年的研究与积累，当年的天才物理学家也已经成为全球复杂网络领域当之无愧的第一人。由于在这一领域做出了卓越的贡献，巴拉巴西更成为诺贝尔奖获奖呼声最高的候选者之一。

一面是全球复杂网络研究第一人，一面是钟爱湖蓝色T恤、只喝健怡可乐的科学狂人



在巴拉巴西提出“无标度网络”的概念之前，科学家惯于将所有复杂网络视为符合泊松分布的随机网络。但巴拉巴西在研究互联网时，却发现互联网并不是随机的。他的实验结果令人非常惊讶：基本上，互联网是由少数链接数多的页面串连起来的，80%以上页面的链接数不到4个。然而只占节点总数不到万分之一的极少数节点，却有1000个以上的链接。这一发现，彻底推翻了“复杂网络是随机的”这一得到人们多年认可的理论，巴拉巴西为复杂网络研究谱写了一个新篇章。

巴拉巴西为复杂网络理论和研究做出了杰出贡献，且贡献涉及社会科学、物理学、数学、计算机科学等多个领域。他让人们第一次真正了解了互联网是如何从最初的一个人类发明变得越来越像一个生命体或生态系统，这背后体现了那些支配所有网络的法则是多么强大。难怪世界著名科技杂志《科技新时代》赞誉道：“他可以控制世界。”

不过，在学术领域之外的日常生活中，巴拉巴西也只是一个普通人。他喜欢过简单的生活，还有坚持自己的小执拗。例如，他唯独钟爱湖蓝色的T恤，所以会一打一打地购买；他喜欢喝可乐，却不接受健怡以外的其他口味；他会利用参会的空闲时间，带上可爱的儿子去丛林里寻找躲起来的老虎；去吃侍者口中“一点儿也不辣”、他们却觉得味蕾要辣到爆炸的食物。



用科学揭示成功定律，用人生诠释成功之路

年少时，物理学让巴拉巴西有机会去探索宇宙和控制人类生命的各种规律。后来，为了寻觅更多的挑战，他开始关注网络与数据中蕴含的复杂性。作为一名警惕的提问者，他接着把好奇心投入到关于“成功”的问题之上——那些挂在现代艺术博物馆里、看起来脏兮兮不起眼的照片，为什么会被认为是杰作？为什么音乐剧《天上人间》而非《猫》，会成为历史上最杰出的音乐剧？去那些学费昂贵的学校就读到底值不值得？为什么每个领域都只有一小撮人能够成为超级巨星？这些问题成了一切的源起。因为钟爱社会结构背后的数学，所以在拿到关于人类成功的海量数据后几年，巴拉巴西终于找到了一种办法将“成功”这个概念分解成若干成分，从而能够像研究螺母和齿轮一样研究它，而且，在人类历史上第一次，揭示出了成功的五大普适定律。

一直以来，巴拉巴西都堪称一位“成功”的科学家，因为他的研究成果极为丰富，而他的文章也是世界顶级科技期刊上的常客。巴拉巴西已经在《科学》和《自然》上发表学术作品30多篇，在《美国科学院院报》上发表作品逾百篇。他的论文被引用总次数接近10万次，H指数高达97，是复杂网络领域被引用最多的科学家。巴拉巴西的研究也获得了各界的高度认可。他现在是美国物理学会院士、匈牙利科学院院士、欧洲科学院院士。巴拉巴西还曾先后获得2005年欧洲生物化学学会联盟（FEBS）颁发的生物系统年度奖项；2006年匈牙利计算机学会颁发的冯·诺伊曼金质奖章；2008年日本C&C基金会颁发的计算机与通信奖、美国国家科学院颁发的美国国家科学奖章；2011年国际工业与应用数学大会颁发的拉格朗日奖。



艾伯特 - 拉斯洛·巴拉巴西相关作品

作者演讲洽谈，请联系

speech@cheerspublishing.com

更多相关资讯，请关注

湛庐文化微信订阅号

湛庐CHEERS 特别制作

译者序
如小说般引人入胜的网络科学

网络科学，是一门关于网络的科学。这门兴起于世纪之交的科学，起源于对现象的观察、对数据的解读、对规律的揭示、对科学普适性的探究，致力于探究自然界广泛存在的网络所具有的一般性规律。网络科学的二十年，是波澜壮阔、跌宕起伏、争议不断的二十年。这样的历程，需要一幅画卷来描绘，本书就是这样一幅画卷；这样一门科学，必将走进教科书，本书就是这样一本教科书。

本书的作者是网络科学奠基人艾伯特 - 拉斯洛·巴拉巴西。作者首先以其个人视角回顾了网络科学的诞生过程，讲述了网络科学从无人问津到炙手可热的发展历程。进而，本书将我们拉回网络科学诞生的20世纪末，用娓娓道来的笔法描绘了当时的时代和科学背景，用提问的方式给出了那个时代的科学家所思考的问题以及思考问题的方式。随后，本书用叙事的方式、科学的语言讲述了图论、随机网络、无标度性质、巴拉巴西 - 阿尔伯特模型、演化网络、度相关性、网络鲁棒性、社区、传播现象等网络科学的内容，抽丝剥茧，深入浅出，在科学和艺术的边界带我们体验网络科学之旅。

本书的主题——网络科学，是一门交叉科学，涉及数学、物理学、社会科学、计算机科学等很多学科。之于数学，毫不为过地说，是随机图论（数学的一个分支）滋养了网络科学。之于物理学，可以说，促使网络科学脱胎于随机图论而独立发展的力量，便来自网络科学的物理学视角——解释真实网络呈现出的普遍现象。之于社会科学，六度分隔、弱链接、小世界等启发网络科学发展的现象大多来自社会科学领域。之于计算机科学，以计算机为代表的信息技术不仅为网络科学提供了强大的研究工具，其快速发展所形成的互联网、万维网、在线社交网络等大规模网络还为网络科学提供了重要的研究对象。相信网络科学这样一个交叉科学，会对很多领域的学生、学者、从业人员有所启发和帮助。

我和网络科学有着很深的渊源。我本人是一名网络科学领域的研究人员。我攻读博士的那段时间，恰好是网络科学在国内快速普及和发展的几年。我的博士学位论文《复杂网络的社区结构》与本书第8章介绍的内容密切相关。目前，我在中国科学院计算技术研究所从事科研工作，带领网络分析和社会计算研究组，长期开展网络科学和网络数据挖掘方面的研究工作。我在中国科学院大学讲授研究生课程《网络数据挖掘》，其中有9个课时的内容是关于网络科学的。这些渊源使我尤其期待有一本关于网络科学的教科书，翻译巴拉巴西撰写的这本《巴拉巴西网络科学》使我夙愿得偿。

2013年，应湛庐文化之邀，在巴拉巴西的《链接》一书出版十周年之际，我重译了这本网络科学早期的重磅之作。新译本在北京发布之时，巴拉巴西出席了发布会，那是我们的第一次见面，既是一次作者和译者的见面，也是一次大师和粉丝的见面。这次见面还促成了不久之后我作为访问学者赴巴拉巴西在美国东北大学的复杂网络研究中心的学术访问。为期一年的访问给我留下了很多美好的回忆。在那里，我近距离地感受了网络科学领跑者们对于网络科学的理解，并了解到一些网络科学诞生过程中的奇闻趣事。这期间，我和巴拉巴西一起合作发表了三篇学术论文，领略了网络科学领军人物开阔的科研视野、炽热的工作热情、高超的写作技巧和平易的人格魅力。在那里，我认识了王大顺、宋朝鸣、贾韬、高建喜、刘洋彧、严钢、池丽萍、靳擎等多位在巴拉巴西实验室从事网络科学研究的中国学生，他们都是我日后在网络科学领域从事科研的好朋友。

在我访问巴拉巴西实验室的那段时间，巴拉巴西完成了这本书的撰写。有一次午餐时间，他谈到希望这本教科书将来可以被翻译成各国文字，我当时就表达了有意翻译这本书的想法。令我有些诧异的是，他不建议我一个科研工作者直接来翻译这本书，而是希望我作为一个组织者联合更多的人完成翻译。于是，在接到湛庐文化的翻译邀请时，我便联系了本书的另一位译者——当时在巴拉巴西的复杂网络研究中心做博士后的黄俊铭博士。俊铭是从我负责的网络分析和社会计算研究组毕业的博士生，我们有很多共同的兴趣爱好和学术合作，于是一拍即合。

具体的分工如下：我负责翻译前言、第1章“图论”、第2章“随机网络”、第3章“无标度性质”、第4章“巴拉巴西 - 阿尔伯特模型”和结语部分；俊铭负责翻译第5章“演化网络”、第6章“度相关性”、第7章“网络鲁棒性”、第8章“社区”和第9章“传播现象”。后来，Facebook的杨旸自愿加入了我们，分担了第4章和第7章的初译，另外我研究组的赵睿卓同学和曹婍同学分别参与了第2章和第3章的初译。翻译结束后，由我来负责整个译本的统稿和校对，以尽量消除翻译语言风格等方面的差异。自问已尽力，但仍不能保证无翻译不当之处，还望读者海涵理解并多提宝贵意见。感谢湛庐文化的安烨编辑以及为本书的排版和出版辛苦工作的同仁，没有你们就没有该译本的顺利面世。感谢我的妻子成玉玉、儿子沈钧昊，感谢俊铭的家人黄炳光、孙文涛、刘丹青、黄湛木，感谢他们对我们翻译工作的理解和支持！

诚如巴拉巴西所言，作为一名科研工作者，我日常的科研工作非常繁忙，直接翻译一本书是一件不太现实、甚至不太划算的事情。我所负责的章节，几乎都是在周末带孩子上课外班时、临睡前、出差的路上等碎片时间完成的。俊铭当时在做博士后，每天有做不完的科研工作和其他事情，在翻译本书期间他的小宝宝诞生了，见证了本书的翻译历程。杨旸的加入，缓解了我们不少的压力，在此对她表示由衷的感谢！最终，我们坚持下来了，我们做到了。

翻译这本书对我而言是一次心灵之旅。特别是在翻译本书的前言部分时，我叹服于巴拉巴西对网络科学的拳拳之心和孜孜追求，这不正是我们每一位科研人员所需要的吗？巴拉巴西在书中讲述的科研挫折和心路历程，引起了我深深的共鸣。当我们处于一个学科诞生的前夕，我们能听到其来临的脚步吗？我们能感受到时代的呼唤并以一种舍我其谁的无畏勇气投入其中吗？巴拉巴西做到了。

本书是一本教科书，有教科书标配的概念、问题、方法、证明、图表、公式、习题。本书是一本巴拉巴西式的教科书，科学感、艺术感、历史感交织辉映。每个章节都有一幅有故事的精美图片作为开篇引言，有趣的案例配有在线视频来阐述，优美细腻的写作方式让这本教科书如小说般引人入胜。

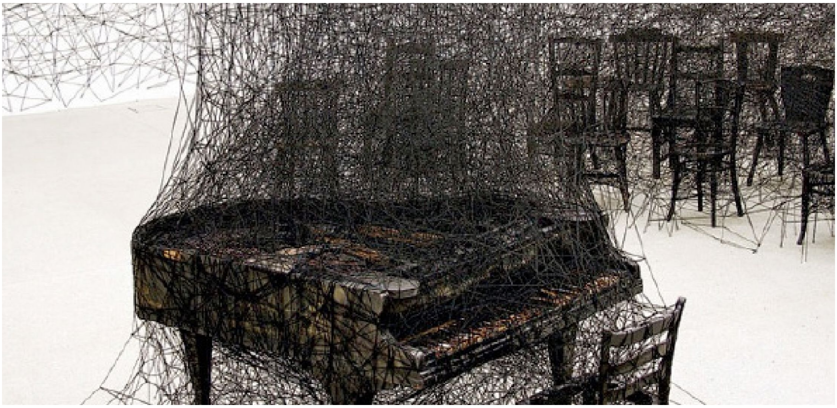
我想对这本书的读者说，若想了解和学习网络科学，这本教科书

是一个很好的选择，阅读这本书会使你和网络科学奠基人巴拉巴西产生一次心灵碰撞。

沈华伟

2019年12月

前言
网络科学诞生的奇妙旅程



图I-0　盐田：在沉默中

盐田是一位在柏林工作的日本装置艺术家。她用羊毛编织的网将普通物品包裹起来。这些网络的随机性包含了有序和无序之间的内在张力，而这正是网络科学的主题。

本章索引

I.1 曲折之路 / **VII**

I.2 参考文献 / **XXV**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森 **SARAH MORRISON**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

阿迈勒·侯赛尼 **AMAL HUSSEINI**

莫罗·马蒂诺 **MAURO MARTINO**

菲利普·赫费尔 **PHILIPP HOEVEL**

妮科尔·萨迈 **NICOLE SAMAY**

I.1
曲折之路

如今，每年有十余个学术会议、研讨会和讲习班在关注网络，有百余本书及多本重点期刊专注于网络领域，大部分大学开设了网络科学课程，有三个大洲的大学设置了网络科学博士学位，科研资助机构也在网络领域投入了数亿美元的经费。看到这些，人们不禁会想，这个仅出现十余年的领域真可谓一帆风顺。不过，受此种效应的影响，我们可能会忽略一个最耐人寻味的问题：网络领域是如何成长得这么快的？

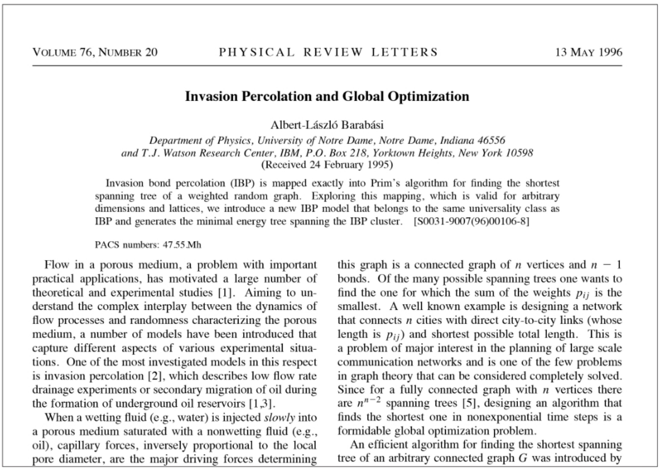
我把这一章视为我的科学发现旅程，原因其实很简单：我并没有打算对上述问题提供一个无偏见的回答，而是希望从一个参与者的角度去回忆网络科学的兴起。我恰好是最了解这个参与者的人，因为“他”就是我。我想回忆的并不是什么胜利大游行，相反，我经历的是一场曲折坎坷的旅行，一场充满挫折、矛盾和冲突的旅行。这一章也不是什么全貌图，而是主要关注我在穿越网络科学的“森林”时，反复碰到的那些令人难以忘记的“树”。需要提醒大家的是，科学发现绝不像教科书中描述的那样直观和自然，而是像接下来几页里你将看到的那样充满曲折。

我的首篇关于网络的论文（1994年）

我对网络的着迷始于1994年，即我在IBM传奇之地——沃森研究中心做短期博士后的那几个月。考虑到假期即将来临——假期期间不能留在IBM，我决定利用这段时间更多地去了解我的雇主（当时，IBM几乎就是计算机的代名词）。于是我去了沃森图书馆，寻找关于计算机科学的入门书籍。

怀着对计算机领域科学问题的好奇心，我从沃森图书馆借阅了一本书。这本书涵盖了很多内容，包括算法、布尔逻辑（Boolean Logic）和NP-完全性（NP-completeness）。我对书中讲述最小生成

树问题的一章特别感兴趣。我发现，书中介绍的克鲁斯卡尔（Kruskal）算法很好地对应了统计物理学领域的著名模型——侵入渗流。于是，在圣诞节过去的两个月后，也就是1995年2月24日，我的第一篇关于网络的论文在《物理评论快报》（*Physical Review Letters*）上发表了[1]。该论文阐述了物理学和计算机科学领域中这两个广泛研究的网络问题之间的等价性（图**I-1**）。在如此有影响力的物理期刊上发表一篇单作者论文对于我的学术生涯来说无疑是重要的，不过这一举措的真正影响远不止于此：这篇论文打开了我对网络科学研究的热情之门，奠定了我随后数十年对网络热爱的基础。



VOLUME 76, NUMBER 20 PHYSICAL REVIEW LETTERS 13 MAY 1996

Invasion Percolation and Global Optimization

Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556
and T.J. Watson Research Center, IBM, P.O. Box 218, Yorktown Heights, New York 10598

(Received 24 February 1995)

Invasion bond percolation (IBP) is mapped exactly into Prim's algorithm for finding the shortest spanning tree of a weighted random graph. Exploring this mapping, which is valid for arbitrary dimensions and lattices, we introduce a new IBP model that belongs to the same universality class as IBP and generates the minimal energy tree spanning the IBP cluster. [S0031-9007(96)00106-8]

PACS numbers: 47.55.Mh

Flow in a porous medium, a problem with important practical applications, has motivated a large number of theoretical and experimental studies [1]. Aiming to understand the complex interplay between the dynamics of flow processes and randomness characterizing the porous medium, a number of models have been introduced that capture different aspects of various experimental situations. One of the most investigated models in this respect is invasion percolation [2], which describes low flow rate drainage experiments or secondary migration of oil during the formation of underground oil reservoirs [1,3].

When a wetting fluid (e.g., water) is injected *slowly* into a porous medium saturated with a nonwetting fluid (e.g., oil), capillary forces, inversely proportional to the local pore diameter, are the major driving forces determining

this graph is a connected graph of n vertices and $n-1$ bonds. Of the many possible spanning trees one wants to find the one for which the sum of the weights p_{ij} is the smallest. A well known example is designing a network that connects n cities with direct city-to-city links (whose length is p_{ij}) and shortest possible total length. This is a problem of major interest in the planning of large scale communication networks and is one of the few problems in graph theory that can be considered completely solved. Since for a fully connected graph with n vertices there are n^{n-2} spanning trees [5], designing an algorithm that finds the shortest one in nonexponential time steps is a formidable global optimization problem.

An efficient algorithm for finding the shortest spanning tree of an arbitrary connected graph G was introduced by

图**I-1　1994—1995**：我在网络领域的首次尝试

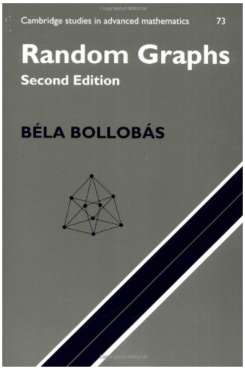
我在1994年的寒假构思出了第一篇关于网络的论文，将计算机科学领域著名的最小生成树问题和统计物理学领域广泛研究的侵入渗流关联在一起。这标志着我对网络科学长期追求的开始。

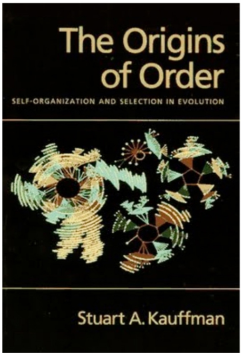
失败1：第二篇论文（1995年）

学得越多，我越觉得自己对真实网络知之甚少。居住在纽约市的我，想象着曼哈顿地下铺设的数百万条电线、电话线和光缆线到底有多么复杂。图论告诉我们，这些网络中的节点是随机连接的，然而我并不这么认为。我坚信，一定有某些组织原则在支配着我们日常生活中使用的网络。寻找这些组织原则恰好是统计物理学家的拿手好戏，

毕竟他们经训练而成的思维方式就游弋在有序和随机的边界。

随后的几个月里，我阅读了贝拉·波罗巴斯（Béla Bollobás）关于图论的优秀著作 [2] ，这使我接触到了保罗·埃尔德什（Paul Erdös）和阿尔弗雷德·雷尼（Alfréd Rényi）在图论方面的经典工作 [3] 。与此同时，斯图尔特·考夫曼（Stuart Kaufmann）富有远见的作品使我意识到网络对于生物领域的重要性 [4] 。在这些书籍中，两种截然不同的视角碰撞在了一起：一边是由定理驱动的枯燥的数学世界，一边是考夫曼那种不受数学羁绊的肆意想象（图**I-2** ）。



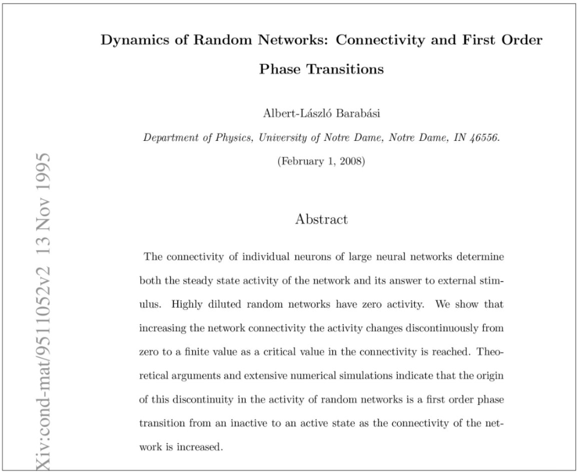


图**I-2**　**1995**：有序，还是随机？

这三本书中的两本启发我开启了网络科学之旅。我现在想不起来第一本书的标题或副标题了，这本我在1994年从IBM沃森研究中心图书馆借阅的书好像是叫“计算机科学的50个问题”。

在博士后岗位上工作了8个月后，我接受了美国圣母大学（University of Notre Dame）提供的教职，这使我在IBM余下的4个

月内可以专注地完成我的第二篇关于网络的论文。这篇题为《随机网络动力学：连通性和一阶相变》[5]的论文，是我探究网络拓扑结构的首次尝试。该论文将波罗巴斯的世界和考夫曼的世界结合在一起，探索了网络结构的改变会如何影响布尔系统的演变（图**I-3**）。论文背后的观察非常简单：改变随机网络的平均度，布尔系统将经历一次相变。因此，如果不能完全理解网络的结构，你就无法解释布尔系统的行为。



Dynamics of Random Networks: Connectivity and First Order Phase Transitions

Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, IN 46556.

(February 1, 2008)

Abstract

The connectivity of individual neurons of large neural networks determine both the steady state activity of the network and its answer to external stimulus. Highly diluted random networks have zero activity. We show that increasing the network connectivity the activity changes discontinuously from zero to a finite value as a critical value in the connectivity is reached. Theoretical arguments and extensive numerical simulations indicate that the origin of this discontinuity in the activity of random networks is a first order phase transition from an inactive to an active state as the connectivity of the network is increased.

Xiv:cond-mat/9511052v2 13 Nov 1995

图**I-3　1995—1997**：一直没有发表的网络论文

这是我在网络科学领域的第二次尝试，也是我探究网络拓扑作用的首篇论文。在被4个期刊拒稿后，这篇论文于1995年11月出现在Arxiv的在线服务器上。我最终放弃了将其发表在期刊上的想法。

这篇论文受到了很多想法的启发，这些想法源于蜂窝网、互联网和万维网。然而，我过去发表论文的物理期刊几乎不涉及此类研究。因此，我努力尝试，去寻找上述发现在我所在领域中的应用。最终，我决定把我的发现应用于神经网络——物理学家经常研究的一个主题。在我看来，关注神经网络的物理学家应该会认可网络的重要性。然而，我错了。这次决定标志着随后4年间我的网络科学之旅中一系列失败的开始。

1995年11月10日，我将完成的论文投稿到《科学》杂志，之后

便回到波士顿参加美国材料研究学会的年度会议。在那里，我碰到了《自然》杂志的编辑菲利普·鲍尔（Philipp Ball），他对交叉学科很感兴趣。我有幸向他介绍了那时令我非常着迷的新研究对象——网络。于是，当我的论文在几周后被《科学》杂志拒稿时，我将它发给了菲利普，希望《自然》杂志会对它感兴趣。的确如此，论文被送审了。

不过，论文的评审人对网络并不那么着迷。其中一位评审人在他的评审意见中坦率地写道：

1. 论文的研究动机不清晰；

2. 技术上很受限；

3. 关于演化和互联网的推测不切实际。

当然，这位评审人是正确的：论文中没有解释为什么我们需要关注网络。答案都在我脑子里，但由于我博士毕业不过一年，仅仅靠4年前读博士时才开始学习的英语，实在是难以把脑中的想法通过文字变成故事。

我很失望，并于1996年4月25日将该论文再次投稿到《物理评论快报》——结果并没有好多少，论文经过漫长的评审过程之后再次被拒收。1997年11月21日，在首次投稿两年后，我又把该论文投稿到《欧洲物理快报》（*Europhysics Letters*）。那时候，我的网络科学之旅的第二个重大失败已经开始了。

失败2：绘制万维网地图

当我为第二篇论文的发表而付出艰苦努力时，我开始感觉到，要想继续前进，就需要放弃我一直使用的图论，而应该做物理学家擅长的事情：从真实世界中寻找灵感。也就是说，我需要地图——确切地说是真实网络的地图。

在蒂姆·伯纳斯－李（Tim Berners-Lee）开放万维网5年后，也就是谷歌创立的两年前，那时的万维网还处于成长期。许多在实验室中开发出来的搜索引擎，譬如JumpStation、RBSE Spider或

Webcrawler，开始尝试绘制万维网链接结构的地图。1996年2月，我向一些运行此类网络爬虫工具的研究人员发送了邮件（图**I-4**　），希望他们能够提供部分数据——当然，最好是有完整的地图，哪怕没有这样的地图，能够知道每个节点拥有多少个链接也足够了。“我希望使用这些数据绘制一个直方图。”我写道。我希望得到这个在三年之后被我们称为“万维网的度分布”的直方图。

Date: Fri, 09 Feb 1996 10:34:17 -0500
From: Albert-Laszlo Barabasi <alb@nd.edu>
To: [redacted].gov
Cc: alb@nd.edu
Subject: Robots
X-Url: [redacted].html

Dear [redacted],

I am doing some research on random networks and their statistical mechanical properties. The best available real word example of such networks is the WWW with its almost random links. To try out my approach I need some data that that Robots could provide without much difficulty. I friend of mine (who knows much more about the dangers of writing and operating a poorly working- robot) convinced me that instead of attempting to write my own robot, I should rather check if somebody with an already running robot could either (i) help me with the data I need or (ii) allow me to use his/her robot for this purpose.

I wonder if you are willing to give me a help in this direction? Of course, any help will be carefully acknowledged when the results of this research will be published (this is all-academic, non-profit basic research).

When a robot visits a new site, it finds a number of external links (pointing to other home pages). I need statistics regarding this number. Robots regularly collect this information, since this is how they assemble their database. Thus the only thing I need is to have the robot write this info into a file in a structured form, that would
allow me to extract this information. Maybe some of the robots do save the obtained data in a format that would allow me to simply collect these numbers.

For example, if the robot visits the home page http://www.new.homepage.edu/bbb.html it finds that there are for example four
links there, pointing to the addresses:
http://www.aps.org/xxx.html
http://www.my.best.friend/home.html
http://www.my.hobby/joke.html
http://www.my.preffered.newspaper/news.html

So the type of list I could most use is this one (or something
equivalent):
http://www.new.homepage.edu/bbb.html
HAS LINKS TO:
http://www.aps.org/xxx.html
http://www.my.best.friend/home.html
http://www.my.hobby/joke.html
http://www.my.preffered.newspaper/news.html

Moreover, to start with it would be enough less information, for
example a just listing the number of links he found:

4

After visiting a fair number of home pages the table would look
like this:

4
2
3
2
0
19
10
1
0
1

How many datapoints do I need? Well, I wish to make a simple histogram of the previous data, thus I need enough data to obtain a smooth histogram. This histogram will be the starting point of my investigation.

I hope you are willing to help me to obtain this information. If you are not running your robot currently, but are willing to lend me your code so that I can run from my computer to collect this data (I have an IBM RISC 6000 that I could use for this purpose), that is also a solution. Again, I do not plan to use the robot for any other purpose than collecting this (and similar) statistics on the connectivity of the web. If you are interested in more details regarding the nature of the scientific questions I am investigating, I am happy to provide it to you.

laszlo

Albert-Laszlo Barabasi
Assistant Professor

图I-4　1996：寻求数据

这是我寻求数据时发出的一封邮件。邮件发给了一些在20世纪90年代搭建网络爬

虫的计算机科学家，我希望他们能够共享一些关于万维网拓扑的数据。回过头来看，这并不是一封非常有说服力的邮件，没有人回复也就不足为奇了。我不得不再等两年，直到郑浩雄（Hawoong Jeong）加入我的实验室并搭建了我们自己的网络爬虫。这个网络爬虫收集到的数据让我们发现了无标度网络。如果能够像我希望的那样，在1996年就得到这些数据，我们可能会提前三年发现无标度网络。

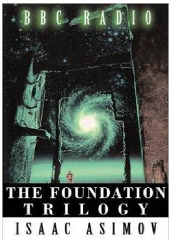
没有人说“不”。实际上，没有人顾得上答复我。就在等待回复的时候，我的第二篇论文遭遇了最后一击——它被《欧洲物理快报》拒稿了。

到那时为止，我的网络之旅可谓非常让人失望。我的第二篇关于网络的论文向4个期刊投稿，经过了三轮评审。没有人说它是错误的，评审人的意见很简单：谁关心网络呢？那时候，我打算获取真实数据的“B计划”也进入了死胡同。我感到失望，同时也承受着发表论文和申请经费的压力，于是，我逐渐开始从网络转向相对稳妥些的量子点方面的研究。

我真的是别无选择。成为助理教授两年后，我的首项经费资助才姗姗来迟，要成为终身教授更是前景渺茫。虽然我是那么相信网络，过去三年里拿得出手的也只有一篇论文和一系列的失败罢了。不过，转向更传统的研究课题让我得到了回报：到1997年年底，我得到了两项经费资助，这使我可以雇用几个学生和一个博士后了。

复兴（1998年）

1997年，我居住在芝加哥，每隔一天去一趟圣母大学。为了打发无聊的两个小时车程，我开始听一些书的磁带。有一天，我在图书馆看到一本童年时读过的书——艾萨克·阿西莫夫的《基地》（图**I-5**）。浏览这本书时，我被哈里·谢顿（Harry Seldon）能够预测数百年后人类命运的能力深深吸引了。



图I-5　1997：复兴

阿西莫夫的科幻小说《基地》三部曲促使我重返网络科学领域。

《基地》是一本极好的科幻小说：非常有趣，描述的未来遥不可及，但从某种抽象角度来看又是合理的。

连接圣母大学和芝加哥市的90号公路两侧是大片的玉米地，这使我能够打开思绪，思考很多荒诞的问题：怎样才能把阿西莫夫的科幻世界变成现实？是否可以设计出一些方程，从而对社会这样复杂的系统进行预测呢？为了达到这个目标，我能做些什么？在我关于量子点的研究进行得有声有色时，阿西莫夫把我的思考拉回到那些让我经历了很多挫折却从未失去兴趣的事物上——网络和复杂系统。

1998年年初，我打算再试一次。我规划了一个与网络相关的新研究项目，于3月在圣母大学最好的餐厅邀请了我的学生雷卡·阿尔伯特（Réka Albert）共进午餐。雷卡那时攻读硕士学位刚满一年半，但成绩斐然。她关于颗粒介质的论文刚刚成了《自然》杂志的封面文章，而她正在进行的研究项目的初步结果看上去也非常好。因此，我邀请她共进午餐的目的看起来很不现实：我希望说服雷卡放弃她几乎就要成功的研究，我想让她和我一起研究网络。

当我让我最好的学生加入这场网络之旅时，我没有什么鼓励的话可以对她讲。我不得不告诉她，我关于网络的第二篇论文被4个期刊拒稿，可能永远都无法发表。网络这个研究题目在当时没有学术群体，没有专业期刊，也没有资助。我必须坦诚地对她说，似乎并没有人关心这个研究题目。因此，从事网络研究有很大的风险，可能会让她之前的成功之路戛然而止。

然而，我同时也告诉雷卡，成功必须经历风险。而且在我看来，

网络值得赌一次。

午餐后，我把一个密密麻麻写满字的文档交给了雷卡，上面写的是我对网络科学的构思。我设想，大概需要6个月时间量化网络拓扑，接下来还需要6个月去理解网络拓扑对网络动力学的影响。然后，我们便可以转向真实的问题，探索网络拓扑和网络动力学的共同演化。

当然，我心里完全没谱：我无法预见拓扑所能提供的重要意义。不过，那不是当时的重点，重要的是她平静而优雅的回应——雷卡同意了，她加入了我前途未卜的网络之旅。

失败3：小世界（1998年）

1999年之前，那些研究网络的科学家几乎毫无交集，时至今日我仍然百思不得其解。一部分原因是，当时存在着一个小而活跃的社会网络学术群体，该群体的形成可以追溯到20世纪40年代。实际上，我们今天关于小世界问题所知道的很多东西来自1960年左右的一篇鲜为人知的论文，该论文的作者是社会学家伊锡尔·德索拉·普尔（Ithiel de Sola Pool）和数学家曼弗雷德·科亨（Manfred Kochen）。虽然这篇论文在1978年之前一直没有发表 [6] ，它的预印版却在社会网络学术群体中流传甚广，并且启发了斯坦利·米尔格拉姆（Stanley Milgram）在1967年完成小世界实验 [7] 。而米尔格拉姆的工作在1/4个世纪后又启发了剧作家约翰·格尔（John Guare）创造出“六度分隔”这个概念。

普尔和科亨所采用的模型，与同一时期的图论学家埃尔德什和雷尼提出的模型一样。但是，在社会学的论文中，没有任何证据表明社会学家知道当时出现的关于随机图的大量数学文献。另外，埃尔德什和雷尼的开创性工作虽然启发了很多随机图方面的工作，但是图论领域没有人知道社会网络学术群体的存在，他们关于小世界的研究成果也没被引用过。

这两个学科之间的界限集中体现在两个学术群体所关注的问题不同：图论学家关心的是相变、子图和巨连通分支，而令社会学家着迷

的则是小世界、弱链接和社区。对社会学家而言，拥有一百个节点的网络便难以理解了，而令数学家兴奋的是节点数趋于无穷大时的极限情况。

1998年，当邓肯·瓦茨（Duncan Watts）和史蒂夫·斯托加茨（Steven Strogatz）关于小世界网络的论文[8]在《自然》杂志上发表时，我首先想到的就是我第二篇关于网络的论文在该杂志上发表失败的尝试。在阅读他们这篇论文时，我发现自己那篇论文失败的原因显而易见，让人恼火：我犯了一个“取景错误”。两篇论文都采用了随机网络范式。但是，我问的问题是物理学家感兴趣的，而论文定位的对象却是神经科学家。相反，瓦茨和斯托加茨问的问题根源于社会学，六度分隔为他们的稿件做了很好的叙述铺垫。

同时，小世界模型对于我和雷卡那时正在研究的问题而言似乎是一个终结。作为物理学家，我们关注的是不能由随机性产生的模式。因此，我们当时正在寻找真实网络中存在而那些固态物理学广泛研究的规则网络和埃尔德什与雷尼的纯随机网络不能刻画的现象。瓦茨-斯托加茨模型介于规则网络和随机网络之间，而规则网络和随机网络正是我们试图规避的两个极端。因此，我们把他们这篇论文放在一边，将其视为行进道路上的干扰。仅仅几个月后，当我们意识到小世界可以为我们的网络之旅提供一些意想不到的帮助时，我再次拿出了这篇论文。

绘制网络地图（1998年）

1998年，当郑浩雄以博士后身份加入我的研究组时，雷卡和我已深深沉浸在网络研究中。毕业于韩国著名大学——首尔大学的郑浩雄有着丰富的计算机方面的知识。1998年秋天的一个晚上，我来到他的办公室，与他讨论他在量子点方面的研究进展——那时候他的主要研究项目是量子点。我们不经意间谈到了网络，这促使我向他讲述了我获取万维网真实拓扑数据时的失败经历。我问他是否知道如何搭建一个网络爬虫——网页采集器的通俗说法。郑浩雄说他之前没有做过，但愿意试一下。就这样，几周后，郑浩雄的网络爬虫就开始忙着爬进

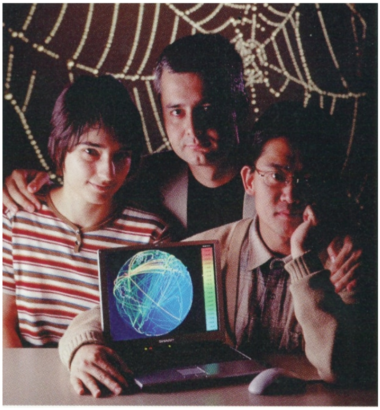
万维网了。我那曾经失败的探索万维网结构的“B计划”复兴了。

我们决定使用郑浩雄收集到的数据来延续我在1996年搁置的研究（图**I-4**　），测量万维网的度分布。我们被一个简单的问题驱动：万维网到达它的逾渗阈值了吗？埃尔德什和雷尼预言，当网络的链接密度小于某个临界值时，网络会破碎成很多个彼此不连通的团。而一旦链接密度达到临界阈值，网络中会涌现出一个巨连通分支，该连通分支大到几乎可以将其视为整个网络。

万维网还能被分裂成许多彼此不连通的部分吗？还是像当时人们所认为的那样已经连成一个大网络了？无论答案是什么，这个问题都非常有趣。为了回答这个问题，我们需要知道万维网的度分布。现在，郑浩雄的网络爬虫可以帮助我们做到这一点。当从郑浩雄收集的数据中绘制出万维网的度分布时，我们着实吃了一惊：我们没有看到随机网络理论所预言的泊松分布，出现在眼前的是一个幂律分布。

郑浩雄的数据展示出的结果和我在过去4年的网络之旅中学到的一切大相径庭。在有关网络的各类文献中，幂律度分布并没有被提及过。实际上，在那个时候，几乎没有人注意到幂律度分布：关于随机图和社会网络的文献都想当然地认为度分布是泊松分布。我们观测到的幂律分布预测，万维网中存在枢纽节点——枢纽节点是拥有大量链接的节点，它们在随机图中几乎不可能出现。当时已有的模型都无法解释枢纽节点的出现。

1999年3月30日是我的32岁生日，那天，我给郑浩雄发了一封邮件，告诉他我要开始撰写我的第三篇网络论文。一个很容易想到的写法是专注于真实发现，这很简单：万维网代表着一种新型的网络，一种之前人们尚未认识到的组织形式。不过，我隐隐感觉到，这样写会是一个错误。那时候，我已经开始认为，我的第二篇网络论文的失败与科学没有多少关系，而是因为前面提到的“取景错误”。如果专注于本质，观测的科学价值就显得过于枯燥，不可能会引起《自然》杂志编辑们的兴趣。于是，我决定尝试特洛伊木马的方式：把这个发现藏在六度分隔的后面。我们给这篇论文起名为“万维网的直径”，声称万维网中的六度分隔实际上是“十九度分隔”（图**I-6**　），然后就把论文寄给了《自然》杂志[9]。



图I-6　1999：“十九度分隔”团队

照片来自2000年的《商业2.0》杂志。照片上的人是雷卡·阿尔伯特、郑浩雄和我，照片是在我们关于万维网拓扑的论文发表不久后拍摄的。

发现（1999年）

论文提交后不久，我赴西班牙和葡萄牙进行了为期两周的长途旅行，旅行的最后一站是一个在波尔图大学召开的研讨会。驾车穿越伊比利亚半岛时，我的脑海里一直萦绕着这样的问题：为什么会有枢纽节点呢？为什么是幂律分布呢？

为了弄清楚是什么使万维网如此特别，我们需要了解更多的网络。因此，在飞往欧洲之前，我开始尝试寻找更多的网络地图。第一幅网络地图来自杰伊·布罗克曼（Jay　Brockman）。布罗克曼是圣母大学的一位计算机科学教授，他向我们提供了一个IBM生产的计算机芯片的布线图。邓肯·瓦茨给了我们一个电网的地图，布雷特·查登（Brett　Tjaden）向我们分享了好莱坞演员数据库。我把这些都留给了雷卡，让她在我旅行这段时间内分析一下这些网络地图。

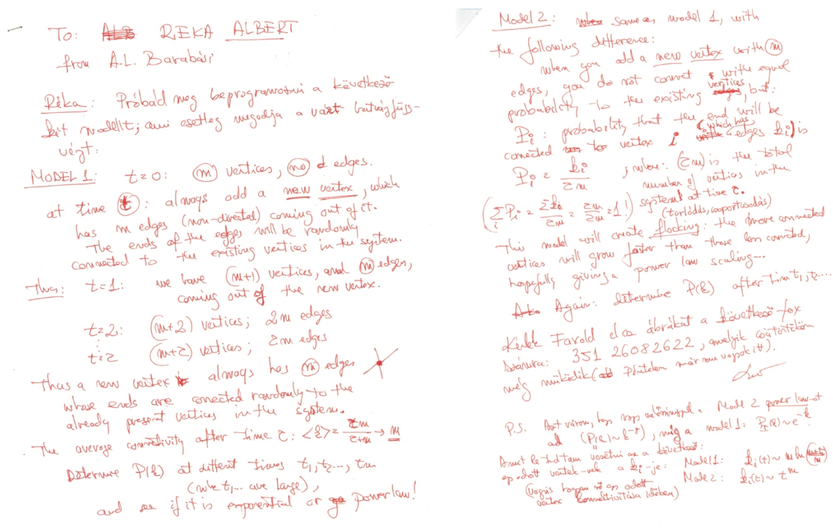
1999年6月14日，雷卡通过电子邮件向我汇报了她的分析结果，那时候我已经在波尔图了。在电子邮件的最后，她写下了这样一行文字：“我查看了这些网络的度分布，在几乎所有这些系统中，度分布的尾部都遵循着幂律分布。”

雷卡的这句话就像晴天霹雳一样。我再也不能专心地听报告了，

脑海里开始不停地思考这个发现的内涵。差异很大的万维网和好莱坞演员网络，居然具有同样的幂律分布，我们之前在万维网上看到的幂律分布居然是普遍的。因此，幂律分布的出现一定是由于某个法则或机制在起作用。而且，既然幂律分布适用于演员网络、计算机芯片网络和万维网等不同类型的网络，关于它如何出现的解释一定是非常简单和基本的。

我需要一个安静的地方来仔细思考。于是，我离开研讨会，回到我在会议期间居住的神学院。

不过，我没有走太远。就在从波尔图大学步行到神学院的15分钟内，我找到了解释。我在房间内开始计算，试图将自己的想法变成数学语言。然后，我发送了一个传真给雷卡，让她做一些数值模拟，以验证我快速想到的这个结论（图I-7 ）。



图I-7 “无标度”传真

这是我于1999年6月14日从葡萄牙波尔图发给雷卡·阿尔伯特的传真。这条匈牙利语和英语混杂的传真描述了如今被称为“巴拉巴西－阿尔伯特模型”的算法。该模型解释了无标度网络形成的原因，并勾勒出了计算度指数的连续介质理论（第4章 ）。

几个小时后，她的回复通过电子邮件发过来了。令我感到吃惊的是，我的结论居然可行。一个非常简单的模型，只含有两个要素——

生长和偏好连接，却很好地解释了我们在万维网和好莱坞演员网络中观测到的幂律分布。

匆忙（1999年）

从葡萄牙回去后，我只能在圣母大学停留7天，然后就要去罗马尼亚的特兰西瓦尼亚（Transylvania）开始为期一个月的假期。然而，我不愿意等到一个月后再将这个发现公开。这就意味着，在葡萄牙的剩余2天加上回美国后的7天里，我就得把论文写好。

我想立刻就开始。不过，我女朋友提醒我，我答应过她在旅行的最后2天不工作的。我们原计划要去里斯本度假。于是，我只好先陪她一起在里斯本游玩，把论文写作计划推迟到从里斯本飞往纽约的8小时的航班上。不过，旅途中，我根本无法将网络从脑海中抹去：当我在圣克鲁斯（Santa Cruz）的狭窄街道上漫步时，论文的雏形逐渐形成了。

飞机一起飞，我就急匆匆地拿出电脑开始打字。在我刚刚写完论文引言部分的时候，正在把一杯可乐递给我身边乘客的空姐一不小心把可乐全洒在了我的键盘上。她不仅毁掉了我那台崭新的笔记本电脑，还打碎了我计划在飞机上完成论文初稿的梦想。

我不能就此放弃。最终，我在那个懊悔不已的空姐向我提供的便笺本上完成了论文初稿。一周后，论文投稿到了《科学》杂志。

不过，我好像是患上了妄想症。那时候，我们有两篇论文在评审中。第一篇论文报告了我们在万维网中的发现，这篇论文正在《自然》杂志的评审过程中。第二篇是刚刚投稿到《科学》杂志那篇，论文指出无标度性质是普适的，并且提出了一种理论来解释无标度性质出现的原因。没错，《自然》和《科学》是最著名的两大学术期刊，但这两个期刊的拒稿率也很高，只有不到10%的投稿最终能够发表。这意味着我们的两篇论文都被拒收的概率超过了81%，而都被接收的概率还不到1%。

我曾经的那篇网络论文在我最终放弃之前，经历了长达两年的坎

坷投稿历程。如果这两篇论文也遭遇了同样的命运该怎么办？如果其他人也同时做出了同样的发现该怎么办？我们发现的这个现象是如此明显、随处可见，有人独立发现它的可能性非常大。我需要一个备选方案。

物理学领域有一个惯例，一篇简短论文在《自然》、《科学》和《物理评论快报》上发表之后，通常紧接着会有一篇长论文发表，以介绍更多的细节。于是，在那7天里，雷卡、郑浩雄和我又写了一篇长论文。我打电话给*Physica A* 期刊的主编吉恩·斯坦利（Gene Stanley），告诉他说，如果他能够快速给出决定，我就打算把自己最炙手可热的论文投稿到*Physica A* 。我不知道他是否相信我所说的“炙手可热”。不过，他答应了会快速评审。

没过几天，我的妄想就被证明是有根据的。我刚刚到达我的家乡特兰西瓦尼亚，就收到了《科学》杂志的拒稿邮件。虽然有些失望，但我坚信这篇论文的结论是重要的。于是，我做了一件之前从没有做过的事：打电话给拒收我论文的编辑，做最后的努力，去尝试改变他的决定。

令我吃惊的是，我成功了，他将我的论文送审了。几个月后，1%的可能变成了现实：《自然》、《科学》和*Physica A* 都接收了我的论文[9]，[10]，[11]。虽然出乎意料，但这是对我过去5年在网络科学领域经历的诸多挫折的最好回报。

180度转变（1999年）

四个学生和一个博士后——那时候我的研究组并不大。除了雷卡之外，其他人都在研究面和量子点。在那篇投稿到《科学》杂志的论文被接收几天后，我召开了一次研究组会议，并在会上做出了一个让研究组成员感到震惊的声明。我告诉他们，我打算退出材料科学的研究。原因很简单：我不想再分散时间和精力。在距离评终身教授还有三年的时候，我决定转变研究领域，从量子点转到网络。我让研究组的成员选择：要么和我一起开始新的旅程，要么离开。

两个学生离开了，其他人选择跟随我一起开启新的未知旅程。

失败4：经费（1999年）

如果我们进入的是一个已经建立起来的领域，并且做好了研究工作，那么获得科研经费只是时间问题。然而，进入一个还不存在的领域将使我们面临不可预知的困难。

好消息是，那时我刚刚获得美国国家科学基金（NSF）在量子点方面的经费支持。坏消息是，我对这个研究题目不再感兴趣。当然，我假装在从事量子点方面的研究，从而继续使用这项经费。不过，这样的做法让我感到内心不安。于是，我打电话给国家科学基金的管理部门，询问是否可以使用这项经费从事网络方面的研究。

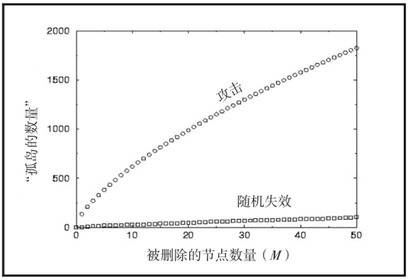
答复是“不行”。要么做量子点方面的研究，要么把经费退还给国家科学基金。这是一个进退两难的境地：一方面，我有经费去从事我已经失去兴趣的增量式研究；另一方面，对于那个有可能会实现突破的研究，我却没有经费支持。

最终，我选择了梦想，把经费退了回去。这令我的研究组面临很大的压力：我们急需新的经费支持，但是还没有资助机构认可网络这样一个研究题目。后来，我注意到美国国防部高级研究计划局（Defense Advanced Research Projects Agency，简称DARPA）在征集“让未来的网络在受到攻击时仍然能够提供服务”的技术。

现在回过头看，这明显是为计算机网络设计方面的专家准备的——计算机网络设计是计算机科学的一个活跃分支。不过，我坚信，没有人能在缺少对网络拓扑及其固有脆弱性的初步理解的情况下构建出具有容错能力的网络。我们刚刚发现的网络无标度性质及随之而来的枢纽节点，对DARPA打算研制的技术一定具有重要意义。有我们在《自然》和《科学》杂志上刚刚发表的论文作为基础，我决定立刻开始撰写申请书。

不过，我们还需要一个确凿的证据，让DARPA的项目管理者相信网络拓扑对于网络健壮性的重要作用。因此，雷卡开始通过从无标度网络中随机删除节点来模拟互联网节点的随机失效。她对比了随机失效和网络攻击（删除枢纽节点）对网络连通性的影响。二者的差异非

常大：无标度网络在随机失效面前表现得非常健壮，但在面临攻击时却异常脆弱。我们立刻把这一发现写进项目申请书，并且相信我们已经毫无疑问地证明了网络拓扑在网络容错能力方面的重要意义（图**I-8**）。



图**I-8　1999**：健壮性

这幅图来自我们于1999年11月1日提交的DARPA申请书，展示了网络拓扑对网络容忍错误和攻击能力的影响。这幅图的原始标题预见了我们一年后发表在《自然》杂志上的那篇关于容错的论文[12]：

"攻击和随机失效对电网的连通性的影响。我们构建了一个有40　000个节点的电网，该网络中一个节点拥有k个链接的概率为$P(k)\sim k^{-3}$。假定攻击一般而言针对系统中拥有链接数最多的节点。为了研究攻击带来的影响，我们从网络中删除链接数最多的M个节点。上面一条曲线展示了'孤岛'数目关于M的函数。例如，删除链接数最多的10个节点，会导致整个网络碎裂成500个'孤岛'——孤岛彼此之间无法通信。随机失效等概率地影响网络中的每个节点，由于大部分节点只拥有少数几个链接，失效很少会带来大的影响。实际上，随机删除M个节点只能让网络分成少数几个不连通的团，而网络的整体连通性基本不受影响。"

11月1日，项目申请书提交了。随后，我对雷卡和郑浩雄说，我们在申请书中提出的问题非常令人兴奋，不能等到DARPA的经费支持批复了再开始研究这个问题，必须立刻行动。于是，在对我们的发现进行扩充之后，我们将其投稿到《科学》杂志。又一次，论文没有送审就被拒收了。

我再次打电话给《科学》杂志的主编，得到的回应是，在他看来，这篇论文的结果与我们之前那篇论文相比没有重大进步。我目瞪口呆，却无法说服他，只好把论文改投到《自然》杂志。

几个月后，DARPA拒绝了我们的项目申请。但是，《自然》杂志

接收了我们的论文，并作为封面论文发布（图I-9）[12]。



图I-9　2000：阿喀琉斯之踵

2000年7月27日那期《自然》杂志的封面。封面上重点介绍了我们关于复杂网络容忍攻击和错误能力方面的论文。这篇论文受到了我们那个（未获资助的）DARPA申请书中的想法启发[12]。

失败5：滑稽的错误

每当有重大科学发现出现时，总会有一些研究人员将保持宇宙平衡作为其人生使命，竭尽所能地将该科学发现从地球上抹掉。网络科学在展现出其革新力量之前，也必须经受命中注定的质疑。

“滑稽的错误”是加州理工学院控制论学家约翰·多伊尔（John Doyle）在其对网络科学十余年的质疑中经常使用的短语，他自称是网络方面的专家。小世界性质是让人惊奇的，它很容易被想到，而且已经有几十年的研究历史了。无标度性质却不同，因此招致了很多我们当时无法回答的质疑：如果这一性质是普适的，那么为什么十余年来一直没有人发现？生长和偏好连接是否只是无标度性质众多解释中的一个？自帕累托时代起，我们就已经知道了幂律，这次又有什么不同呢？贝拉·波罗巴斯则更直接，我们在布达城堡第一次见面时，他对我说，如果没有严格的数学证明，无标度性质可以说“根本不存在”。

这些质疑是有原因的：只有经历过数学训练的研究人员才能理解无标度的概念。由此导致的理解真空给误解和讹传提供了空间。约翰·多伊尔那个言过其实的说法正好填补了这个空间——每当有记者将这

个问题抛给他时他总会这样说。

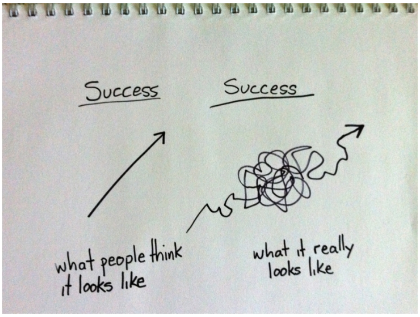
接下来，风向慢慢变了。何塞·门德斯（José Mendes）、谢尔盖·多罗戈夫茨夫（Sergey Dorogovtsev）和锡德·雷德纳（Sid Redner）使用速率方程方法从连续介质理论的角度为无标度网络提供了严格的数学基础[13]，[14]。贝拉·波罗巴斯和他的一些合作者在一篇里程碑式的论文中为无标度性质提供了严格的证明[15]。什洛莫·哈夫林（Shlomo Havlin）和他的学生将网络健壮性和逾渗理论关联在了一起[16]，而贝拉·波罗巴斯和奥利弗·赖尔登（Oliver Riordan）为之提供了严格的证明[17]。一系列随之而来的发现揭示了无标度性质如何深刻地改变了网络的行为，譬如罗穆阿尔多·帕斯托尔-萨托拉斯（Romualdo Pastor-Satorras）和亚历山德罗·维斯皮那尼（Alessandro Vespignani）关于无标度网络上不存在疾病传播临界值的经典发现[18]。学术界由此开始意识到度分布在网络研究中的核心地位。接下来我们将在本书中看到，几乎所有网络特性的解读都必须考虑度分布，包括六度分隔、网络健壮性和社区结构。关于网络的许多根本性问题的提出吸引了数以百计的研究人员从事网络研究，网络科学因此逐渐形成。

小结

如果愿意，你也可以将前面提到的这些事情看作一系列的成功：在接下来的十余年中，1999年《科学》杂志发表的那篇关于无标度网络的论文成了物理科学中引用次数最多的论文。2000年《自然》杂志发表的那篇关于网络容忍失效和攻击方面的论文不仅是该期刊的封面文章，而且对于我们理解网络健壮性产生了深远影响。雷卡和我在接下来的一年又写了一篇关于网络的综述论文，详细介绍了这个领域的理论基础，该论文是《现代物理评论》（*Review of Modern Physics*）引用次数最多的论文[19]。美国科学院2015年出版的美国国家研究委员会（National Research Council）年度报告中使用了“网络科学”一词，并建议美国政府投入数亿美元的经费支持这个新领域的研究，同时将该领域视为一个新的独立学科。最后，两家著名的科学出版社——剑桥大学出版社、牛津大学出版社，一个顶级的工程学会——美

国电气电子工程师学会（IEEE），各自创办了一个期刊，专注于网络科学领域的研究进展。从各个角度来看，我们都可以说一个新的学科诞生了，在一个学科高度交叉的充满活力的学术群体的支持下诞生了。

科学的成功之路从来都不是一条直线（图**I-10** ）。新思想的产生往往需要数年的酝酿。你可以把无标度网络理论视为一个例外，这个思想的火花从想法产生到论文投稿只用了10天时间。然而，若没有那5年毫无成果的磨炼，这个火花也不会燃烧成大火。



图**I-10** 通往成功的不同路径

一位卡通画家绘制的关于成功的作品。该作品生动地刻画了我早期从事网络科学研究的曲折历程，其中充满了各种失败和死胡同。

网络科学提醒我们注意科研合作和导师制度在科学中的重要作用。在雷卡和郑浩雄加入我的网络科学之旅之前，我所拥有的只是一系列的想法和失败。如果没有郑浩雄的网络爬虫技术，我们根本不可能发现无标度性质。雷卡的数学能力为我们探究无标度模型背后的理论提供了保障。如果不是美国西北大学的医师和研究人员佐尔坦·欧尔特沃伊（Zoltán Oltvai）说服我们将网络应用到细胞生物学中，并耐心地指导我们去熟悉那奇妙的蛋白质和代谢物，我们随后在生物网络领域的研究将不会发生[20]，[21]，[22]。这些发现不是一个人完成的，而是真正的共同发现。

如今，许多领域将网络科学归于它们的范畴。数学家声称网络科学是图论的一部分；社会学家认为网络科学属于他们研究了几十年的社会网络范畴；物理学则为网络科学的研究提供了普适性概念和许多

解析工具；生物学领域投入了数亿美元的经费绘制亚细胞网络；计算机科学从计算机算法角度为研究大规模网络提供了帮助；工程领域则在基础设施网络方面投入了大量精力。难以想象，这么多完全不同的学科是如何走到一起，共同哺育了这个新学科的。

这本教科书见证了网络科学诞生的奇妙旅程，以及其中取得的一连串令人兴奋的进展。网络科学后续的成功依赖于我们保持其多学科特性的能力，我们希望每个科学家都能为网络科学的发展带来自己独有的视角。这种思想和观点的碰撞正是网络科学的强大之处和力量之源。

I.2 参考文献

[1] A.-L. Barabási. Invasion Percolation and Global Optimization. *Physical Review Letters* , 76:3750, 1996.

[2] B. Bollobás. *Random Graphs* . Cambridge University Press, 2001.

[3] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)* , 6:290-297, 1959.

[4] S. Kaufmann. *Origins of Order: Self-Organization and Selection in Evolution* . Oxford University Press, 1993.

[5] A.-L. Barabási. Dynamics of Random Networks:Connectivity and First Order Phase Transitions. http://arxiv.org/abs/cond-mat/9511052, 1995.

[6] I. de Sola Pool, M. Kochen. Contacts and influence. *Social Networks* , 1:5-51, 1978.

[7] S. Milgram. The Small World Problem. *Psychology Today* , 2: 60-67, 1967.

[8] D.J. Watts and S.H. Strogatz. Collective dynamics of‘small-

world' networks. *Nature* , 393: 409-10, 1998.

[9] H. Jeong, R. Albert and A.L. Barabási. Internet: Diameter of the world- wide web. *Nature* , 401:130-131, 1999.

[10] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[11] A.-L. Barabási, H. Jeong, R. Albert. Mean-field theory for scale free random networks. *Physica A* , 272:173-187, 1999.

[12] R. Albert, H. Jeong, A.-L. Barabási. Error and attack tolerance of complex networks. *Nature* , 406: 378-482, 2000.

[13] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking, *Physical Review Letters* , 85: 4633, 2000.

[14] P. L. Krapivsky and S. Redner. Statistics of changes in lead node in connectivity-driven networks, *Physical Review Letters* , 89:258703, 2002.

[15] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms* , 18:279-290, 2001.

[16] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. Resilience of the Internet to random breakdowns. *Physical Review Letters* , 85:4626, 2000.

[17] B. Bollobás and O. Riordan. Robustness and Vulnerability of Scale-Free Random Graphs. *Internet Mathematics* , 1, 2003.

[18] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters* , 86:3200-3203, 2001.

[19] R. Albert and A.-L. Barabási. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics* , 74: 47, 2002.

[20] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L.

Barabási. The large-scale organization of metabolic networks. *Nature* , 407:651-655, 2000.

[21] H. Jeong, S. P. Mason, A.-L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature* , 411:41-42, 2001.

[22] A.-L. Barabási, and Z.N. Oltvai. Network biology:understanding the cell's functional organization. *Nature Reviews Genetics* , 5:101-113, 2004.

目 录

从被拒稿的天才物理学家到诺贝尔奖热门人选
一面是全球复杂网络研究第一人，一面是钟爱湖蓝色T恤、只喝健怡可乐的科学狂人
用科学揭示成功定律，用人生诠释成功之路
译者序　如小说般引人入胜的网络科学
前言　网络科学诞生的奇妙旅程
引言　复杂科学的未来
第1章　图论
第2章　随机网络
第3章　无标度性质
第4章　巴拉巴西 - 阿尔伯特模型
第5章　演化网络
第6章　度相关性
第7章　网络鲁棒性
第8章　社区
第9章　传播现象
结语　运用网络科学理解当今的互联世界

引言
复杂科学的未来

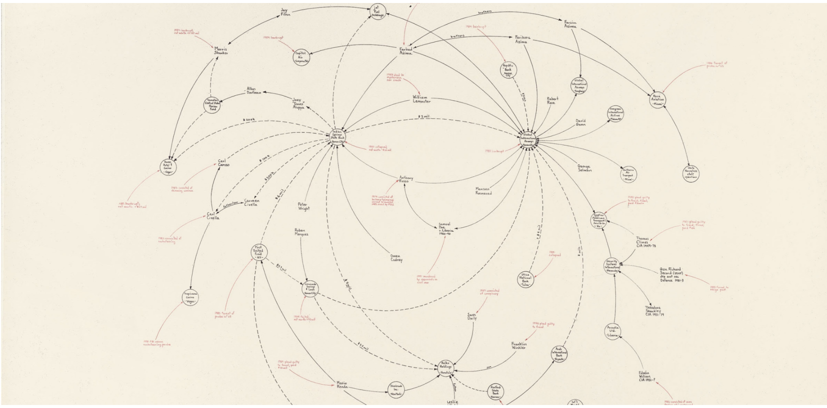


图**0-0**　马克·隆巴尔迪：全球国际航空和印度斯普林州立银行

马克·隆巴尔迪（Mark Lombardi，1951—2000）是一位美国艺术家，致力于描绘“权力的使用和滥用”。在他开展该领域的研究工作之前，已经有很多详细的研究存在了，其数量之多超出了他的应对能力。因此，隆巴尔迪尝试把它们组合成手工绘制的关系图，以便他能更专注于自己的工作。后来，这些关系图成了一种艺术[1]。此图采用了1977年到1983年间兴起的一种画法，该画法使用彩色铅笔和石墨进行绘图。

本章索引

0.1 互联带来的脆弱性 / **003**

0.2 复杂系统的核心——网络 / **005**

0.3 推动网络科学出现的两股力量 / **007**

0.4 网络科学的特性 / **011**

0.5 社会影响 / **013**

0.6 科学影响 / **019**

0.7 小结 / **024**

0.8 课后习题 / **026**

0.9 参考文献 / **027**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森 **SARAH MORRISON**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

阿迈勒·侯赛尼 **AMAL HUSSEINI**

莫罗·马蒂诺 **MAURO MARTINO**

菲利普·赫费尔 **PHILIPP HOEVEL**

罗伯塔·西纳特拉 **ROBERTA SINATRA**

0.1
互联带来的脆弱性

乍一看，图**0-1** 中的两幅卫星图像并没有什么区别：光芒四射的区域是人口密集区，黑暗笼罩的地方则是无人居住的森林和海洋。然而，仔细观察之后，我们便能看出其中的差异：图a中闪耀着光芒的多伦多、底特律、克利夫兰、哥伦布和长岛在图b中消失在了黑暗中。这并不是下一部“世界末日”电影里的场景，而是2003年8月14日前后美国东北部的真实情景：一场大面积停电波及了美国东北部8个州、约4500万人，以及安大略湖区的1000万人。

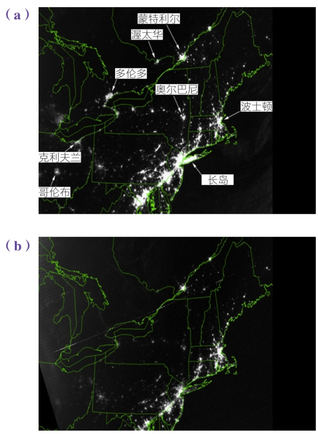


图**0-1** **2003**年北美大范围停电

（**a**） 2003年8月13日美国东部时间晚上9点29分拍摄的美国东北部各州的卫星图像，距离2003年北美大范围停电发生还有20个小时。

（**b**）停电发生5小时后的卫星图像。

2003年这次大面积停电事件是级联失效的典型案例。网络就像一个运输系统，一个地方的局部失效会使其原本的负载转移到其他节点上。当这些额外负载比较小时，系统可以轻松地将其吸收掉，局部失效被无形化解。然而，当失效节点转移给其邻居节点的额外负载过多时，其邻居节点也会随之失效，负载相应地转移给它们的邻居。一瞬

间，级联失效就发生了。级联失效的影响大小取决于最初失效节点在网络中的位置及其容量。

级联失效在很多复杂系统中发生过。互联网中，为躲避失效路由器而对通信流量重新路由时，级联失效便会发生。正常的路由操作有时会造成拒绝服务攻击，导致正常运转的路由器因负载过大而失效。我们经历过金融系统中的级联失效。1997年，国际货币基金组织向几个太平洋国家的中央银行施压，促使其下调贷款额度，多家大型企业的资金链因此断裂，最终引起全球范围的股灾。2009年至2011年期间的金融危机经常被当作级联失效的经典案例——美国次贷危机导致全球经济瘫痪，大量银行和企业倒闭，甚至有政府宣布破产。级联失效也可能是人为导致的。例如，为削弱恐怖组织的运转能力，全球政府共同努力使其资金供应枯竭。与之类似，癌症研究人员试图通过触发细胞的级联失效来杀死癌细胞。

美国东北部停电事件展示了本书的几个重要主题：首先，为避免破坏性的级联失效，对于承载级联传播的网络，我们必须理解其结构特征；其次，我们必须能够对网络上发生的动态过程进行建模，例如电流；再次，我们还需要揭示网络结构和其上发生的动态过程之间的互动是如何影响整个系统的健壮性的。尽管级联失效看上去是随机且不可预测的，实际上它遵循着可再现的规律，这些规律可以被量化，我们甚至可以通过网络科学工具对其进行预测。

停电事件还展示了一个更大的主题：互联带来的脆弱性。实际上，在电力投入应用的前些年，每个城市有各自独立的发电站和电网。然而，电力难以储存，一旦生产出来就必须立刻被消费掉。从经济角度考虑，相邻的城市间开始互联，从而可以共享剩余产能，并在需要时相互借电。今天，电的低价格要感谢电网——把所有电力生产者和消费者连接在一起的网络。电网使低成本生产出的电力能够立刻传输到任何地方，为我们提供了一个网络给人类生活带来正面影响的鲜活例子。

网络的各个部分对于网络整体的影响都不容忽视。例如俄亥俄州某地发生保险丝熔断这样的局部失效，其影响或许不只局限于当地。局部失效的影响可以沿着网络的链接扩散出去，影响到看似不相关的

其他节点、消费者和个体。总体而言，互联带来了非局部性，使信息、流行语、商业举措、电力、能源和病毒在各自的社会技术网络上扩散出去，无论距离多远，无处不及。网络在带给我们便利的同时也带来了脆弱性。揭示那些提升网络便利的因素固然很好，但除此之外，本书的主要目标之一是限制那些使网络脆弱的因素。

0.2 复杂系统的核心——网络

> 在我看来，下一世纪将是复杂性的世纪。
>
> ——斯蒂芬·霍金

边栏0.1

复杂

1. 由很多彼此连接的部分构成的；组合的；复合的。例如：复杂的高速公路系统。
2. 由许多繁杂的部件、单元等构成的。例如：复杂的机械。
3. 非常困难的、难以理解的。例如：复杂的问题。

我们生活在许多复杂系统的包围中。例如，人类社会需要数十亿人的合作；通信基础设施连接着数十亿的手机、计算机和卫星；我们理解和认识世界的能力需要大脑中数十亿神经元的协同工作；我们的生命存在体现为成千上万个基因和细胞代谢体之间的紧密交互。

这些系统统称为复杂系统。称其为复杂系统是因为我们难以从对其各个组成部分的理解中推断出系统的整体行为。复杂系统在我们的日常生活、科学、经济中扮演着重要角色，因此对复杂系统的理解、数学描述、预测和控制是21世纪的主要科学挑战之一。

网络科学在21世纪伊始的出现生动地表明了科学可以应对复杂性

的挑战。实际上，在每个复杂系统背后都有一个错综复杂的网络，网络刻画了复杂系统各个组成部分之间的交互关系：

（1） 刻画基因、蛋白质和代谢体之间交互关系的网络，把各个部分有机地组织成鲜活的细胞。细胞网络的存在是生命的必要条件。

（2） 刻画神经元之间连接关系的神经网络，是我们理解大脑运转方式和人类意识的关键。

（3） 职场关系、朋友关系和家庭关系的总和通常被称为社会网络。社会网络是人类社会的重要组成部分，决定着知识、行为和资源在社会中的传播。

（4） 描述通信设备之间通过电缆或无线彼此交互的通信网络，是现代通信系统的核心。

（5） 由发电机和传输线路构成的电网为几乎所有现代技术提供着能源。

（6） 人们彼此交换物品和服务而形成的贸易网络，是第二次世界大战以来世界物质繁荣的基础（图**0-2**）。



图**0-2** 经济系统中的网络

在大英博物馆推出的“百件文物看世界史”展览中，图中的信用卡被选为其中第99件文物。这张信用卡形象地展示了现代经济的高度互联特性——这种彼此间紧密的经济和社会联系在过去一直未被关注。

2009年，这张卡由总部位于伦敦的汇丰银行（HSBC）在阿联酋签发。这张信用卡的使用遵循总部位于美国的信用卡联盟提供的VISA协议，以及受伊斯兰银行（Islamic Bank）管辖。

网络也是21世纪一些革命性技术的核心，催生了谷歌、Facebook、思科（CISCO）和Twitter等大型互联网公司。最后，网络对于科学、技术、商业和自然的意义，要比我们想象的大得多。总之，如果缺少对复杂系统背后的网络的深刻理解，我们就永远不可能理解复杂系统。

21世纪的第一个10年间，人们对网络科学的浓厚兴趣源于这样一个发现：虽然复杂系统多种多样，但这些系统背后的网络，无论是结构还是演化方面，都由一些共同的基本规律和原理支配着。因此，尽管真实世界的网络在形态、大小、性质、年龄和范畴上千差万别，它们都遵循着同样的组织原则。如果不考虑网络中每个组成部分和链接的具体特性，我们所得到的网络彼此都非常相像。在随后的章节中，我们将探讨促使网络科学这一新研究领域出现的推动力及其对科学、技术和社会的影响。

0.3
推动网络科学出现的两股力量

网络科学是一个新学科。对于该学科的精确起始时间，人们或许有不同的见解。不过，大家都一致认为，网络科学作为一个独立的学科是在21世纪出现的。

为什么网络科学没有在200年前出现呢？毕竟该学科所研究的多数网络都不是什么新事物：代谢网络可以追溯到生命起源，有大约40亿年的历史了；社会网络则和我们人类一样历史悠久。而且，包括生物化学、社会学、脑科学在内的很多学科，几十年来一直在研究自己的网络。作为一个数学成果丰硕的子领域，图论自从1735年就开始研究网络了。综上所述，我们有充分的理由把网络科学称为21世纪的科学吗？

在21世纪伊始，一些特别的事情发生了，这些事情突破了单个研究领域的范畴，催生了网络科学这一新学科的出现（图**0-3**）。要想理解为什么网络科学出现在今天而不是200年前，我们需要对推动网

络科学出现的两股力量进行探讨。

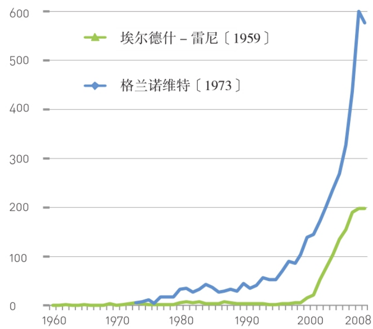


图0-3　网络科学的出现

虽然关于网络的研究有着悠久的历史——起源于图论和社会学，网络科学的新篇章却是在21世纪的第一个10年间才出现的。

人们对网络的兴趣日渐浓厚，可以从对网络科学的两篇经典论文的引用模式中看出。第一篇是保罗·埃尔德什和阿尔弗雷德·雷尼于1959年发表的论文 [2] ，该论文标志着图论领域开始研究随机网络；第二篇是马克·格兰诺维特（Mark　Granovetter）于1973年发表的论文 [3] ，该论文是社会网络领域中引用次数最多的论文。图中展示了这两篇论文自发表以来每年获得的引用数。两篇论文在它们各自的领域中都获得了高度认可，但在各自领域之外的影响却很有限。21世纪以来，网络科学的出现使得这两篇论文的引用数迅速增长，这得益于网络科学给它们带来了全新的跨学科的关注。

网络地图的出现

对于由数百到数十亿个彼此连接的组成部分构成的系统，要想详细描述其行为，我们需要一幅描绘该系统连接模式的地图。对于社会系统，我们需要一个列表，上面准确列出每个人的朋友、朋友的朋友等。对于万维网，我们需要一幅地图来标明网页之间的连接关系。对于细胞，我们需要的地图是一个列表，上面详细列出基因、蛋白质和代谢体之间的相互作用和化学反应。

在过去，我们缺少描绘网络的工具，要记录下刻画网络的大量数据也很困难。互联网技术革命为我们提供了快速而有效的数据共享方法和廉价的数字存储方式，从根本上提高了我们收集、整理、共享和分析真实网络数据的能力。

得益于技术进步，我们在千禧年来临的时候见证了网络地图的爆炸式出现（边栏**0.2**　）。例如，CAIDA和DIMES项目绘制出第一个大规模的互联网地图；生物学家们投入数亿美元，通过实验绘制出人类细胞中的蛋白质相互作用网络；Facebook、Twitter、领英（LinkedIn）等社交网络公司在为准确获取人们之间的朋友关系和职场关系努力着；美国国家卫生研究院（NIH）的Connectome项目致力于系统地描绘出哺乳动物大脑的神经元连接关系。到了20世纪末，大量涌现的网络地图加快了网络科学的出现。

边栏**0.2**

网络地图起源

如今网络科学家们研究的少数几个网络地图是为了研究网络而生成的。大部分网络原本只是其他一些研究项目的副产品，在网络科学家的手中才变成了网络地图。

（**1**）　细胞中的各个化学反应，是生物化学家们历经150多年逐个发现的。到了20世纪90年代，这些化学反应被收集到一个中心数据库时，才提供了将这些化学反应组织成细胞生化网络的契机。

（**2**）　每部电影的演员列表最初是散布在报纸、书刊和百科全书中的。随着互联网的出现，为满足影迷的好奇心，这些数据被收集到了中央数据库，例如imdb.com。这些数据库使得网络科学家们可以从中构建出好莱坞的演员关系网络。

（**3**）　数百万学术论文的作者信息最初也散布在数千种期刊的目录中。最近，科学地图（Web of Science）、谷歌学术和其他学术服务系统将这些信息组织成了数据库，使网络科学家们可以重构出准确的科学合作关系网络。

网络科学早期的历史，很大程度上依赖于研究人员的精心设计，从那些已经存在的数据库中抽取出网络。网络科学改变了这一

切：如今，获得大量资助的联合研究致力于绘制网络地图，以准确刻画生物、通信和社会系统的内部连接关系图。

网络特性的普适性

我们很容易列出在自然或社会中所遇到的不同网络之间的差异性：代谢网络中的节点是微小的分子，节点间的链接是由化学和量子力学定律支配的化学反应；万维网中的节点是网页，链接是供计算机算法解析地址使用的统一资源定位符（URL）；社会网络中的节点是人，链接则表示家庭、职场、朋友和熟识关系。

这些网络的产生过程差异也很大：代谢网络是经过数十亿年的演化而形成的；万维网是由数百万的个人和组织共同努力建成的；社会网络是由数千年来的社会规范塑造而成的。考虑到这些网络所对应的系统在大小、性质、范畴、历史和演化方面有着千差万别，我们对于它们之间存在巨大差异并不感到奇怪。

网络科学的一个关键发现是：这些在不同的科学、自然、技术领域中形成的网络，其架构是彼此相似的，这说明它们在同样的组织原理的支配下形成。因此，我们可以采用同样的数学工具来探索这些网络及其所描述的系统。

普适性是本书的指导原则之一。我们不仅追求对具体网络性质的探索，对于每个具体的网络性质，我们也会探讨其适用范围有多广。我们旨在理解这些性质的形成原因，揭示指导网络演化的规律及其对于网络行为的意义。

综上所述，尽管很多学科对于网络科学都有重要贡献，但这个新学科的出现在相当大的程度上要归功于日益丰富的数据——这些来自各个学科的数据为我们提供了准确的网络地图。进而，这些网络地图使得研究网络科学的科学家们能够从众多网络特性中找出具有普适性的性质。普适性为新诞生的网络科学奠定了基础。

0.4

网络科学的特性

网络科学的内涵不仅取决于其研究对象，更多地还取决于其研究方法。本节，我们将讨论网络科学在理解复杂系统时所用方法的特性。

多学科特性

网络科学为我们提供了一种语言，使多个学科可以借助该语言无障碍地相互交流。实际上，细胞生物学家、脑科学家（图**0-4**　）和计算机科学家面临着类似的任务——描绘复杂系统内部的连接模式，从不完整和有噪声的数据集中抽取信息，理解系统在面临失效或攻击时的健壮性。



图**0-4**　绘制人类大脑

网络科学炙手可热的一个应用领域是关于大脑的研究。对于被称为秀丽隐杆线虫（C.　elegans）的蛔虫，我们很早就获得了其完整的神经系统连接图。不过，我们仍然缺少更大动物的神经连接数据。这一境况如今得到了改变：科学家通过大量努力研制出了可以绘制大脑神经连接图的新技术。2014年4月10日的《自然》杂志封面展示了由西雅图艾伦研究所（Allen　Institute）的研究人员获得的实验室小白鼠的大脑神经连接图[4]。

没错，每个学科都会有不同的研究目标、技术手段和学科自身重

要的科学挑战。然而，这些学科探索的很多问题实际上具有共同的性质，这为跨学科的工具和思想的探索提供了土壤。例如，社会网络领域在20世纪70年代提出的“介数中心度”概念，如今在识别互联网中的高通信负载节点方面发挥着关键作用。与之类似，计算机科学家们为图划分问题开发的算法后来被应用于医药领域的疾病模块识别和大规模社会网络的社区识别。

数据驱动的实证特性

网络科学的一些关键概念都起源于图论——一个高产的数学分支。网络科学区别于图论的地方是它的实证特性——其关注焦点是数据、功能和效用。在随后的章节中，我们将看到，网络科学从不满足于开发描述某个网络性质的抽象数学工具。我们开发的每个数学工具都必须使用真实数据进行测试，然后我们会依据它在揭示系统性质和行为方面提供的洞察力来判断其价值。

定量和数学特性

要想对网络科学的发展有所贡献，或打算正确使用网络科学的工具，就必须掌握其背后的数学形式化原理。网络科学借鉴了图论中的形式化方法来研究网络，借鉴了统计物理学的概念化框架来应对随机性和追求普适的指导原则。另外，网络科学从工程学中借鉴了包括控制论和信息论在内的概念，用于理解网络的控制原理；还从统计学那里借鉴了从不完整和有噪声的数据集中抽取信息的方法。

网络分析软件的开发可以使更多领域的人使用网络科学的分析工具，就连对网络科学的理论基础和数学概念不熟悉的人也可以使用。不过，如果打算更深入地了解网络科学并高效地使用其工具，我们则需要掌握其数学形式化原理。

计算特性

现实中，人们感兴趣的网络往往很大，用以描述它们巨大的数据

规模。网络科学家们在分析这些网络时通常面临一系列艰巨的计算挑战。因此，网络科学具有很强的计算特性，经常需要从算法、数据库管理和数据挖掘领域借鉴方法。有许多软件工具可以解决这些计算问题，帮助具有不同计算技能的研究人员对其感兴趣的网络进行分析。

总之，掌握网络科学需要熟悉该领域上述每个方面的特性。正是这些特性的组合为我们提供了理解真实网络性质所需的多方面的工具和视角。

0.5 社会影响

对一个新的研究领域而言，其影响力可以从科学价值和社会影响两个方面来衡量。社会影响通常是指该研究领域的应用范围和潜在价值。网络科学虽然是一个新学科，其影响却是无处不在的。

经济：从网络搜索到社交网络

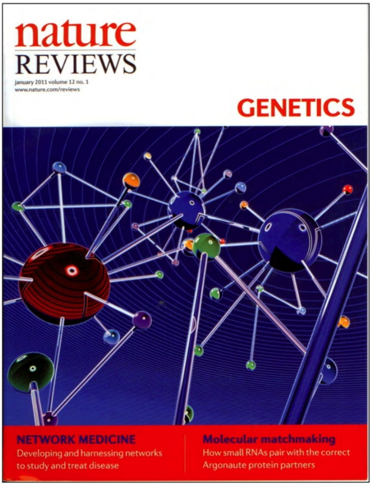
谷歌、Facebook、Twitter、领英、思科、苹果和阿卡迈（Akamai）等21世纪最成功的公司，其技术和商业模式都是基于网络的。实际上，谷歌不仅运营着人类有史以来建造的最大网络——包罗万象且持续更新的万维网，其搜索技术也和万维网的网络特性密不可分。

Facebook的出现大大推动了网络的普及，该公司致力于绘制出全世界人们的社会网络。Facebook不是第一家社交网络网站，很可能也不是最后一家。在社交网络的生态系统中，Twitter和领英等公司仍在努力获取数百万用户的关注。网络科学家们构思出的算法为这些社交网络站点提供了燃料，为包括好友推荐和广告投放在内的几乎所有事情提供了辅助支持。

健康：从药物设计到代谢工程

2001年完成的人类基因组计划首次完整地绘制出了人类基因图谱[5]，[6]。然而，要想完全理解细胞的运转方式和疾病的起源，列出人类所有的基因还不够：我们还需要一幅描绘基因、蛋白质、代谢体及其他细胞组成部分之间相互作用的地图。实际上，消化食物、感知环境变化等大多数细胞过程都依赖于细胞内的分子网络。这些网络的损坏会导致人类疾病。

随着对分子网络重要性的关注日益增多，致力于理解细胞网络行为的网络生物学，作为生物学的一个新的子领域出现了。与之相呼应，医药领域出现了网络医药学，致力于揭示网络在人类疾病中的作用（图**0-5**）。这些进展的重要性可由下面这条消息生动地说明：2012年，哈佛大学成立了网络医药系，公开聘用应用基于网络的思想理解人类疾病的研究人员和医疗人员。



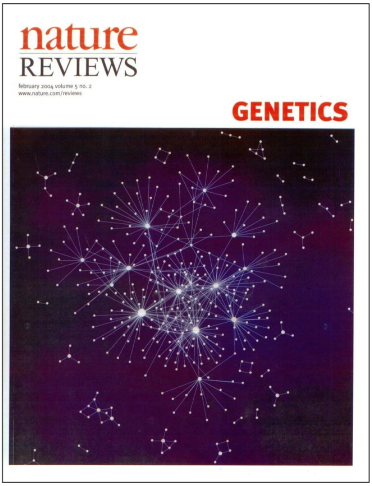


图0-5　网络生物学和网络医药学

遗传学顶级综述期刊《自然综述：遗传学》的两期封面。该期刊特别关注网络的影响：2004年这期封面关注网络生物学（上图）[8]，2011年这期封面探讨网络医药学（下图）[9]。

网络在药物研制方面起着十分重要的作用。网络药理学的终极目标[7]是研制出没有明显副作用就能治愈疾病的药物。为实现该目标，人们在从多个方面努力：投资数百万美元用于绘制细胞网络，开发工具和数据库用于存储、管理、分析病人数据和基因数据。

一些新公司充分利用了网络为健康和医药领域提供的发展机遇。例如，GeneGo从科技文献中收集细胞交互地图，Genomatica利用代谢网络的预测能力识别细菌和人类的药物靶向。最近，强生等大型制药公司也开始大力投资网络医药，将其视为未来药物的方向。

安全：打击恐怖主义

恐怖主义是21世纪的灾难，需要投入大量资源在全世界范围内对其进行打击。负责打击恐怖活动的机构越来越重视网络思维。网络思维被用于破坏恐怖组织的金融网络，绘制敌对关系网络，发现恐怖组织成员的角色和揭示恐怖组织的能力。虽然这个领域的工作大多是机密的，但也有一些公开报道的案例。例如，使用社会网络搜寻萨达姆·侯赛因 [10] ；通过手机通话网络搜寻2004年3月11日马德里火车爆炸案嫌疑人。网络概念还对军事理论产生了影响，形成了“网络中心战”的概念——指通过低强度冲突去对抗采用分布式网络组织的恐怖分子和犯罪网络 [11] （图**0-6** ）。

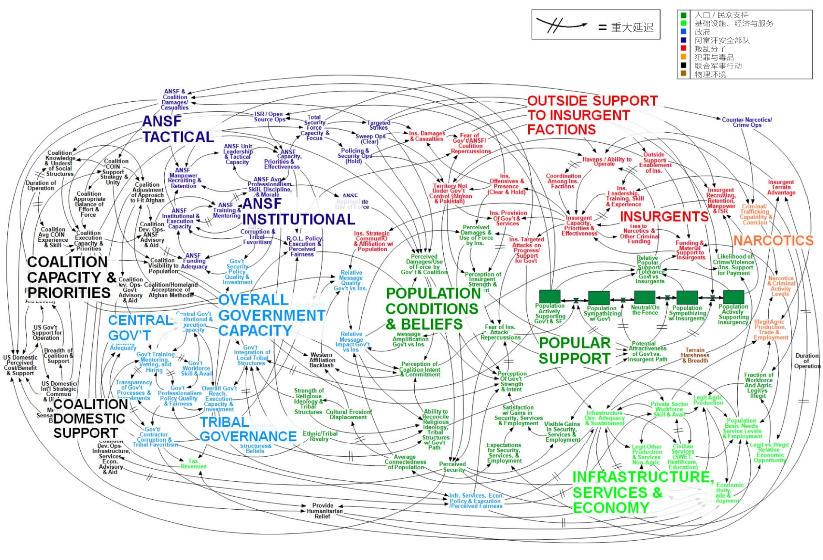


图**0-6** 军事行动背后的网络

这幅图是在2012年阿富汗战争期间设计的，用于描绘美军在阿富汗的部署规划。由于在一个画面中展示了太多的复杂性和细节，这幅图时而成为人们嘲讽的对象，但

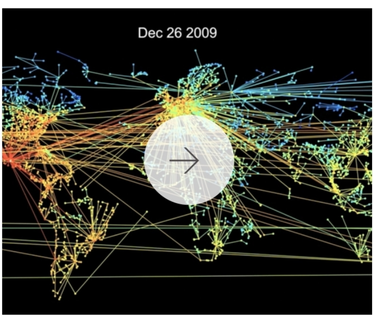
是它生动地刻画了现代军事行动的互联特性。如今，这个例子被军事学院的教官和学生学习，用于阐述网络模型在决策和业务协调中的重要作用。实际上，军事长官的工作不仅仅局限于保障必要的军事能力，还需要考虑当地人的信仰、生活条件等因素的影响。该图摘自《纽约时报》。

网络科学有如此多的潜在军事应用，因此第一个网络科学课程开设在美国军事学院（西点军校）也就不足为奇了。另外，2009年成立的陆军研究实验室（Army Research Lab）投入了超过3亿美元，用于支持全美国的网络科学研究中心。

网络为人们带来的知识和能力也会被滥用，例如美国国家安全局不加区分的网络监视措施[12]。以对抗未来恐怖主义攻击为托词，美国国家安全局对美国和其他国家数亿人的通信进行监视，从而获取他们之间的社会关系网络。通过这个例子，网络科学家们唤醒了一种新的社会责任：确保人们的工具和知识不被滥用。

传染病：从预测到对抗致命病毒

虽然禽流感（H1N1）在2009年暴发时并不像人们担心的那样破坏性巨大，但它在传染病的历史上扮演着一个特别的角色：在暴发高峰到达数月前，其传播过程和传播时间就被准确预测到了，从这个角度而言，禽流感是第一个被预测且被成功拦截的案例（在线资源**0.1**）[13]。人们能做到这一点，主要得益于对疾病传播网络作用的理解。



在线资源**0.1**　预测**H1N1**传染病

2009年对H1N1病毒传播范围的预测，使人类第一次成功实现了对传染病进行实时

预测 [13] 。该项目根据全球交通网络的结构和演化，预言H1N1的传播峰值将出现在2009年10月，而之前人们预计峰值会在下一年的1月、2月间出现。这意味着，那些免疫效果预计在2009年11月出现的疫苗就显得晚了些，对抑制H1N1的传播作用有限。该项目的成功表明，网络科学能够推动人类社会重要领域的进步。视频由亚历山德罗·维斯皮那尼提供。

扫描二维码下载“湛庐阅读”App，回复书名，获取关于传染病传播预测的精彩视频资料。

2000年之前，传染病模型主要是基于区划的模型，假设处于同一个社会物理区划中的所有人都可以互相传染。基于网络的框架带来了根本性的改变，将传染病的可预测性提升到一个新高度。如今，传染病预测是网络科学最活跃的应用领域之一 [13] ， [14] ，例如预测H1N1病毒和埃博拉病毒的传播。基于网络的框架也是本书讨论的一些重要发现的源头，它使我们可以建模并预测生物病毒、数字病毒和社会病毒（例如流行语）的传播。

这些进展的影响超出了传染病的范畴。实际上，网络科学工具在2010年1月预言了病毒可以通过手机进行传播 [15] 。就在预言之后不久，2010年秋天，大规模手机病毒第一次在中国暴发，每天有超过30万部手机被感染。

神经科学：绘制大脑地图

从网络科学的角度来看，由数千亿个相互连接的神经元构成的人类大脑，是目前我们理解最少的网络。原因很简单：我们缺少关于神经元之间连接情况的地图。目前可供研究使用的完全绘制出的大脑只有线虫的大脑，线虫大脑仅包含302个神经元。详细绘制哺乳动物的大脑会带来脑科学的革命，使我们能够理解并治疗很多神经系统疾病和脑疾病。有了大脑地图，大脑研究就会成为网络科学成果最丰硕的应用领域之一 [16] 。受大脑地图潜在影响力的驱动，美国国家卫生研

究院在2010年启动了Connectome计划，旨在研制能够绘制哺乳动物大脑神经元级地图的技术（图**0-4** ）。

管理：揭示组织机构的内部结构

虽然管理主要依靠正式的行政管理系统，但人们逐渐意识到，人与人之间通过交流关系形成的非正式网络对于一个组织的成功起着重要的作用。准确绘制出组织内部的关系网络可以暴露出关键单元之间潜在的交互缺失，有助于识别出那些将不同部门和产品关联起来的重要人物，帮助发现组织内部存在的管理问题。另外，在管理学文献中，越来越多的证据表明，员工的工作效率是由其在非正式组织关系网络中的位置决定的[17] 。

因此，Maven 7、Activate Networks、Orgnet等许多公司都提供工具和方法来绘制出组织的真实结构。这些公司提供的系列服务包括：识别意见领袖，减少员工流失，优化知识和产品的扩散，从多样性、大小和技能方面为具体的任务设计最高效的团队（图**0-7** ）。IBM、SAP等企业在其日常业务中增加了社交网络能力。总之，网络科学的工具对管理和商业来说是必不可少的，可以提高生产率，促进组织内部的创新。

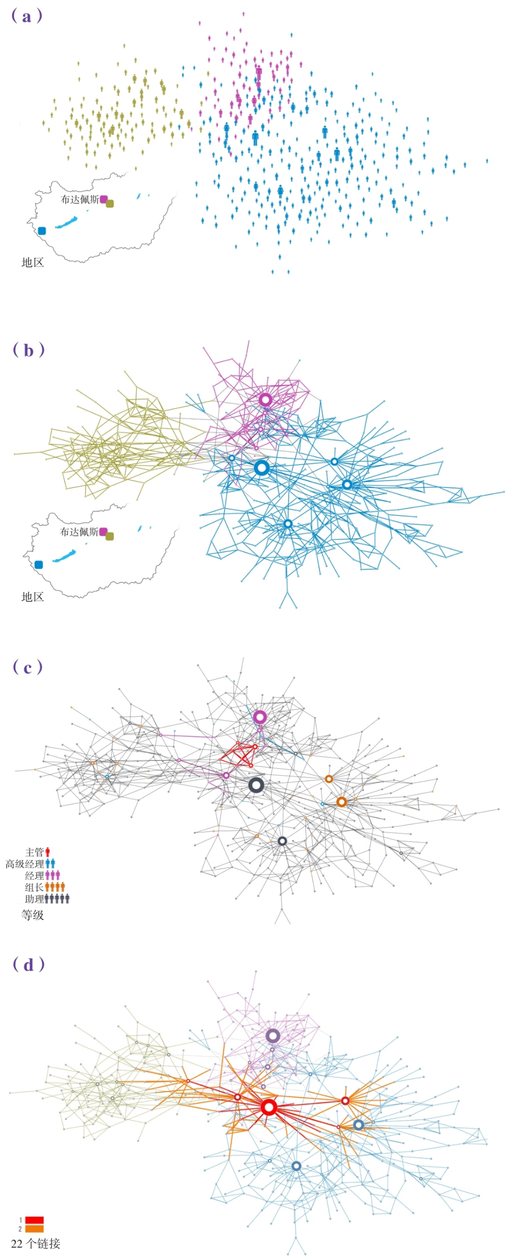


图0-7　组织结构图

（a）　这家匈牙利公司的员工分散在三个办公地点（紫色、黄色和蓝色）。管理部门意识到，体现上级管理部门意图的信息传递给员工之后就变了，看上去和

上级管理部门的真实计划没什么关系了。为了提高公司内部的信息流动，公司管理部门求助于使用网络科学进行组织管理的Maven 7公司。

（**b**） Maven 7开发了一个在线平台，让每个员工写出他们在做和公司相关的决定时会向谁征求建议。该平台得到了如图b所示的网络：如果一个人把另一个人作为征求意见的信息源，则在二者之间连一条线。图中大的枢纽节点对应着高影响力的个体。

（**c**） 个体在公司关系网络中的位置体现了其领导力，颜色表示他们在公司的级别。可以看出，没有一个主管（红色节点）是枢纽节点。也没有高级经理（蓝色节点）是枢纽节点。枢纽节点大多级别较低：经理、组长和助理。最大的枢纽节点即最有影响力的人，是一位普通员工（图中处于中心位置的灰色节点）。

（**d**） 最大枢纽节点的直接链接（红色）和两跳链接（橙色）表明，很大一部分员工距离该枢纽节点只有两跳。然而，这个枢纽节点是谁呢？他是一位负责安全和保洁的员工。他定期会到公司的每个地方去，和那里的员工交谈。除了高层管理人员，几乎其他所有员工都和他有联系。他不了解高层管理部门的真实意图，却能把自己在和员工们交谈中收集的信息传递下去，有效地扮演着流言中心的作用。

公司是应该解雇这个最大的“枢纽节点”，还是给他升职呢？这个问题的最佳解决方案是什么？

0.6 科学影响

网络科学影响最深远的领域还是科学界。《自然》《科学》《细胞》《美国科学院院报》等科学界最著名的学术期刊都刊登了有关网络科学的综述和评论文章，介绍网络给生物学、社会科学等各个学术领域带来的影响。例如，《科学》杂志曾出版过关于网络的专辑，以此纪念无标度网络被发现十周年[18]（图**0-8**）。

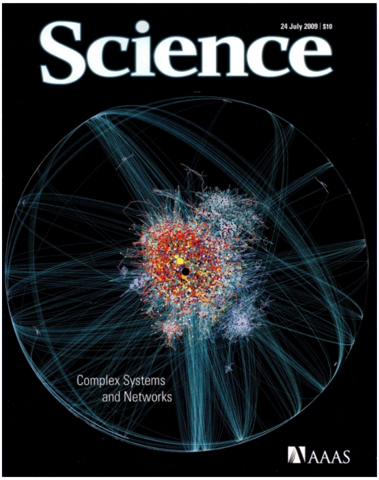


图0-8　复杂系统和网络

为纪念无标度网络被发现（1999年）十周年，《科学》杂志于2009年7月24日出版的网络专辑。

过去10年间，每年都有10余个主题为网络科学的国际会议、研讨会、暑期学习班和冬季学习班。创始于2005年的网络科学旗舰会议“NetSci”是其中一个非常成功的网络科学会议，每年的NetSci会议汇集了全球网络科学领域的学者和从业人员。网络科学的科普书籍成了很多国家的畅销书，推动了公众对网络科学的认识。很多著名大学开设了网络科学课程，学生群体的专业背景非常多样化。2014年，位于波士顿的美国东北大学和位于布达佩斯的中欧大学开始设立网络科学博士学位。

网络对科学界的影响之大，从复杂系统领域引用次数最多的几篇论文中可见一斑。这些论文的单篇引用次数介于2000和5000之间，每篇都是经典论文，涉及的科学发现包括蝴蝶效应、重整化群、自旋玻璃、分形和神经网络。为了展示网络科学的学术影响，我们将前述这些经典论文和两篇引用次数最多的网络科学论文进行了对比（图**0-9**）。这两篇网络科学论文分别是1998年发表在《自然》杂志上的关于小世界现象的论文[19]和1999年发表在《科学》杂志上的报道无标度网络发现的论文[18]。从图**0-9**中可以看出，这两篇论文每年引用数增长速度之快是复杂系统领域前所未有的。

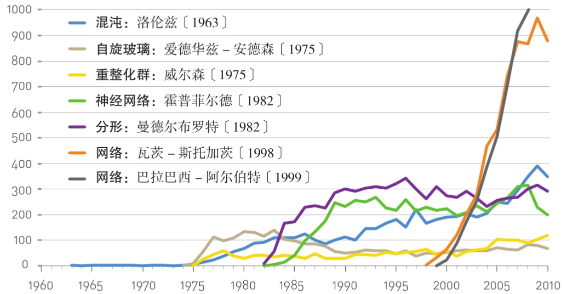


图0-9　复杂性和网络科学

通过和复杂性领域中引用次数最多的论文对比，我们从引用情况来看网络科学的科学影响。20世纪六七十年代，复杂系统研究由三篇论文主导，内容分别是爱德华·洛伦兹（Edward Lorenz）1963年在混沌方面的经典工作 [20] 、肯尼斯·威尔森（Kenneth G. Wilson）提出的重整化群 [21] 、塞缪尔·爱德华兹（Samuel F. Edwards）在自旋玻璃方面的工作 [22] 。到了20世纪80年代，贝努瓦·曼德尔布罗特（Benoit Mandelbrot）发表了关于分形的著作 [23] ，托马斯·威滕（Thomas Witten）和莱恩·桑德尔（Len Sander）提出了受限扩散凝聚模型 [24] ，复杂系统领域的研究焦点随之转移到了模式形成方面。同时期具有同样影响力的工作还有：约翰·霍普菲尔德（John Hopfield）在神经网络方面的工作 [25] ，佩尔·巴克（Per Bak）、汤超（Chao Tang）和库尔特·维森菲尔德（Kurt Wiesenfeld）在自组织临界性方面的工作 [26] 。这些论文在不断刷新人们对复杂系统的认识。图中对比了上述这些里程碑论文及两篇引用次数最高的网络科学论文每年的引用次数。这两篇网络科学论文分别是瓦茨和斯托加茨关于小世界网络的论文，以及巴拉巴西和阿尔伯特发现无标度网络的论文。

从其他角度同样可以看出网络科学对很多学科的重要影响。例如，网络科学的论文成了许多研究领域顶级期刊最常引用的论文：

（1）　根据汤森路透（Thompson-Reuters）科学引文数据库显示，瓦茨和斯托加茨1998年发表在《自然》杂志上的关于小世界现象的论文 [19] 、巴拉巴西和雷卡·阿尔伯特1999年发表在《科学》杂志上的关于无标度网络的论文 [18] ，发表10年后的引用次数均排在物理学领域的前10名。截至2011年，瓦茨和斯托加茨那篇论文是《自然》杂志1998年发表的所有论文中引用次数第二多的论文，而巴拉巴西和阿尔伯特那篇论文是《科学》杂志1999年发表的所有论文中引用次数最多的论文。

（2）　马克·纽曼（Mark Newman）的一篇网络科学论文 [27] 在

《工业和应用数学学会评论》（*SIAM Review*）上发表4年后，成为工业和应用数学学会（Society of Industrial & Applied Mathematics，SIAM）旗下所有期刊上发表的论文中引用次数最多的一篇。

（**3**）《现代物理评论》自1929年创刊以来就是所有物理期刊中影响因子最高的期刊。2012年之前，该期刊上发表过的所有论文中引用次数最多的论文是诺贝尔奖获得者苏布拉马尼扬·钱德拉塞卡（Subrahmanyan Chandrasekhar）于1944年发表的经典综述论文《物理学和天文学中的随机问题》（*Stochastic Problems in Physics and Astronomy*）[28]。发表70年后，该论文累计获得了高达5000多次引用。然而，在2012年，该论文引用次数的领先位置被2001年发表的第一篇网络科学综述论文《复杂网络的统计力学》（*Statistical Mechanics of Complex Networks*）取代[29]。

（**4**）2001年，帕斯托尔－萨托拉斯和维斯皮那尼在《物理评论快报》上发表论文，发现无标度网络中不存在传播阈值[30]。该论文和一篇关于量子计算的论文是《物理评论快报》2001年发表的所有论文中引用次数最多的两篇。

（**5**）米歇尔·格文（Michelle Girvan）和马克·纽曼发表的关于网络社区识别的论文[31]，在《美国科学院院报》2002年发表的所有论文中引用次数最多。

（**6**）2004年发表的题为《网络生物学》（*Network Biology*）的综述文章[8]是遗传学顶级综述期刊《自然综述：遗传学》发表过的所有论文中引用次数第二多的论文。

科学界对网络科学非同寻常的热情，促使负责为美国政府提供政策建议的美国科学院的分支机构国家研究委员会开始关注和考察网络科学。国家研究委员会召集了两个小组，经讨论形成了两份政策建议报告[32]，[33]，界定了网络科学这个领域（图**0-10**）。这两份报告不仅记录了一个新研究领域的出现，还强调了其对科学界、国家竞争力和国家安全的重要意义。在这两份报告发布之后，美国国家科学基金

会成立了网络科学理事会，美国的一些大学在陆军研究实验室的资助下成立了网络科学中心。

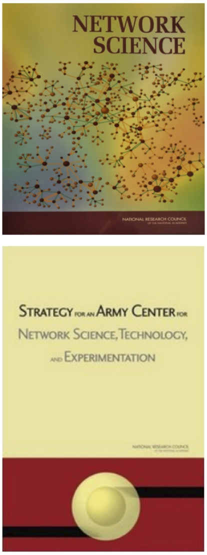


图**0-10**　美国国家研究委员会发布的两份关于网络科学的报告

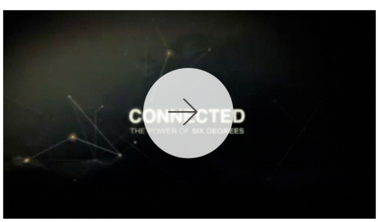
这两份报告记录了网络科学的出现，并指出这门新学科对科研和国家竞争力的长期影响。报告中建议对该领域进行大力支持，推动了美国各所大学建立网络科学中心，并推动了美国国家科学基金会设立网络科学理事会。

公众对网络科学也是青睐有加。这得益于几本科普书的热卖，例如《链接》（*Linked*　）、《联结》（*Nexus*　）、《六度分隔》（*Six Degrees*　）和《大连接》（*Connected*　）（图**0-11**　）。澳大利亚制片人安娜玛丽亚·塔拉斯的一部纪录片将网络科学带上了荧屏，在世界各地播放，并且获得了很多著名的奖项（在线资源**0.2**　）。



图0-11　广泛影响

这4本被译成20多种语言的畅销书将网络科学在公众中进行了普及 [35] ， [36] ， [37] ， [38] 。



在线资源0.2　《连接》

这部由安娜玛丽亚·塔拉斯（Annamaria　Talas）导演的名为《连接》（*Connected*）的获奖纪录片，向我们介绍了网络科学。演员凯文·贝肯（Kevin　Bacon）和一些著名

的网络科学家也因其而出名。

扫描二维码下载“湛庐阅读”App，回复书名，获取关于网络科学的纪录片。

网络科学也给艺术家们带来了启发。有很多与网络相关的艺术作品面世，而且艺术家和网络数据科学家们每年会召开一次联合研讨会[34]。在《社交网络》《六度分隔》等一系列与网络科学相关的科幻小说和故事片的推动下，网络科学已经在流行文化中深深扎下了根。

0.7 小结

网络科学的出现看上去是一个突发现象，实际上该领域是随着网络的作用和重要性日益获得社会广泛认可而出现的（图**0-3**和图**0-9**）。图**0-12**展示了“网络”、“进化”和“量子”这三个词在过去两个世纪中的使用频次。“进化”是谈及达尔文进化论时最常使用的词，“量子”是谈及量子力学时最常使用的词，而进化论和量子力学在过去两个世纪中引发了两次重要的科技革命。如我们预期的那样，“进化”一词的使用频次在1859年达尔文的著作《物种起源》出版之后开始增加；1902年首次出现的“量子”一词，在20世纪20年代之前几乎很少被使用，后来，随着量子力学逐渐被物理学家接受并获得大众的关注，“量子”一词才开始被频繁使用。

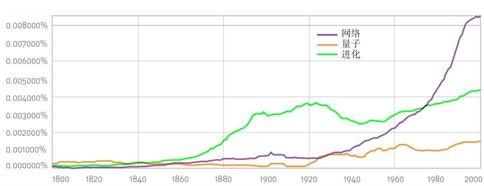


图**0-12** “网络”的兴起

“进化”、“量子”和“网络”这三个词在1800年之后的书籍中被使用的频次。从图中可以看出，“网络”一词在20世纪后期迎来了广泛的社会认知，为网络科学的出现打下了基础。此图由谷歌的n元词组工具生成，刻画了每年使用“进化”“量子”“网络”的书籍比例。

如图**0-12** 所示，“网络”一词的使用频次在20世纪80年代之后开始增加，并超过了“进化”和“量子”。与“进化”和“量子”一样，“网络”一词使用频次的大幅增长体现了网络概念逐渐被社会认可。

进化论、量子力学和网络科学的影响有些类似：它们不仅是拥有核心科学贡献和系统知识体系的科学领域，还是支持很多其他学科发展的使能平台。实际上，遗传学革命就是建立在进化论基础之上的，量子力学为化学、电子学等很多当代科学的进步提供了平台。同样，网络科学也是一个平台，为社交网络、药物设计等很多科学问题的探索提供了新的工具和视角。

由于网络超乎寻常的科学和社会影响力，我们亟须掌握研究和量化网络的工具。本书随后的内容将围绕这一重要主题展开。

0.8 课后习题

习题1　无处不在的网络

列出三个不同的真实网络，陈述每个网络中节点和链接的内涵。

习题2　兴趣分享

说出你本人最感兴趣的网络，并回答下面的问题：

（1）其中的节点和链接分别是什么？

（2）该网络有多大？

（3）该网络可以绘制出来吗？

（4）你为什么对它感兴趣？

习题3　影响

在你看来，网络科学在未来10年会对哪些领域产生最大影响？并给出理由。

0.9 参考文献

[1] J. Richards, R. Hobbs. *Mark Lombardi: Global Networks* .Independent Curators International, New York, 2003.

[2] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae* , 6: 290, 1959.

[3] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology* , 78: 1360, 1973.

[4] S.W. Oh et.al. A mesoscale connectome of the mouse brain. *Nature* , 508: 207-214, 2014.

[5] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* , 409: 6822, 2001.

[6] J. C. Venter et al. The Sequence of the Human Genome. *Science* , 291: 1304, 2001.

[7] A. L. Hopkins, Network Pharmacology. *Nature Biotechnology* , 25: 1110-1111, 2007.

[8] Z. N. Oltvai and A.-L. Barabási. Network Biology:Understanding the cell's functional organization. *Nature Reviews Genetics* , 5: 101, 2004.

[9] N. Gulbahce, A.-L. Barabási, and J. Loscalzo. Network

medicine: A network-based approach to human disease. *Nature Reviews Genetics* , 12: 56, 2011.

[10] C. Wilson. Searching for Saddam: A five-part series on how the US military used social networking to capture the Iraqi dictator. 2010. www. slate.com/id/2245228/.

[11] J. Arquilla and D. Ronfeldt. *Networks and Netwars: The Future of Terror, Crime, and Militancy* . RAND: Santa Monica, CA, 2001.

[12] A.L. Barabási, Scientists must spearhead ethical use of big data. Politico.com, September 30, 2013.

[13] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J.J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck,V. Colizza, and A. Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine* , 7: 45, 2009.

[14] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *PNAS* , 101: 15124, 2004.

[15] P. Wang, M. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science* , 324: 1071, 2009.

[16] O. Sporns, G. Tononi, and R. Kötter. The Human Connectome: A Structural Description of the Human Brain. *PLOS Computional Biology* , 1: 4, 2005.

[17] L. Wu, B. N. Waber, S. Aral, E. Brynjolfsson, and A. Pentland. Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task. Proceedings of the International Conference on Information

Systems, Paris, France, December 14-17, 2008.

[18] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286: 509, 1999.

[19] D. J. Watts and S .H. Strogatz. Collective dynamics of‘small-world’ networks. *Nature* , 393: 440, 1998.

[20] E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences* , 20: 130, 1963.

[21] K. G. Wilson. The renormalization group: Critical phenomena and the Kondo problem. *Reviews of Modern Physics* , 47: 773, 1975.

[22] S. F. Edwards and P. W. Anderson. Theory of Spin Glasses. *Journal of Physics* , F 5: 965, 1975.

[23] B. B. Mandelbrot. *The Fractal Geometry of Nature* . W.H. Freeman and Company. 1982.

[24] T. Witten, Jr. and L. M. Sander. Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon. *Physical Review Letters* , 47: 1400, 1981.

[25] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *PNAS* , 79: 2554, 1982.

[26] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of $1/f$ noise. *Physical Review Letters* , 59: 4, 1987.

[27] M. E. J. Newman. The structure and function of complex networks. *SIAM Review* . 45: 167, 2003.

[28] S. Chandrasekhar. Stochastic Problems in Physics and Astronomy. *Reviews Modern Physics* , 15: 1, 1943.

[29] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks. *Reviews Modern Physics* , 74: 47, 2002.

[30] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters* , 86:3200, 2001.

[31] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS* , 99: 7821, 2002.

[32] National Research Council. *Network Science* .Washington, DC: The National Academies Press, 2005.

[33] National Research Council. *Strategy for an Army Center for Network Science, Technology, and Experimentation* .Washington, DC: The National Academies Press, 2007.

[34] M. Schich, R. Malina, and I. Meirelles (Editors). *Arts,Humanities, and Complex Networks* [Kindle Edition], 2012.

[35] A.-L. Barabási. *Linked: The New Science of Networks* .Perseus Books Group, 2002.

[36] M. Buchanan. *Nexus: Small Worlds and the Groundbreaking Science of Networks* . Norton, 2003.

[37] D. Watts. *Six Degrees: The Science of a Connected Age* .Norton, 2004.

[38] N. Christakis and J. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives* .Back Bay Books, 2011.

第1章 图论

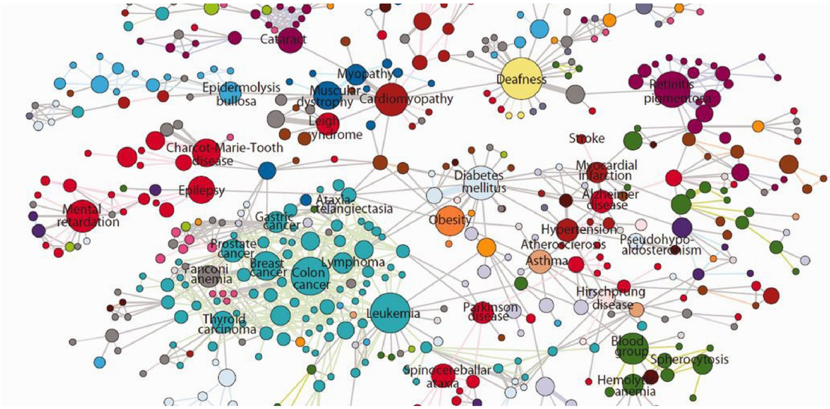


图1-0　人类疾病网络

在人类疾病网络中，节点表示疾病，两个节点间的链接表示两种疾病有共同的基因起源。该网络发表在《美国科学院院报》的增刊上 [1] ，用于展示不同疾病之间的基因关联。随着时间推移，该网络跨越了学科边界，开始作为一个独立的生命而存在。《纽约时报》构建了该网络的交互版，位于伦敦的蛇形画廊（Serpentine Gallery）——世界顶级的当代艺术画廊之一。蛇形画廊在关于网络的篇章中展出了该网络 [2] ，另外，它还出现在许多关于设计和绘图的书籍中 [3] ，[4] ，[5] 。

本章索引

1.1 哥尼斯堡的桥 / **033**

1.2 网络和图 / **035**

1.3 度、平均度和度分布 / **038**

1.4 邻接矩阵 / **042**

1.5 真实网络是稀疏的 / **044**

1.6 加权网络 / **046**

1.7 二分网络 / **048**

1.8 路径和距离 / **051**

1.9 连通性 / **056**

1.10 集聚系数 / **058**

1.11 小结 / **060**

1.12 课后习题 / **063**

1.13 进阶阅读**1.A**
全局集聚系数 / **066**

1.14 参考文献 / **068**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森 **SARAH MORRISON**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

莫罗·马蒂诺　**MAURO MARTINO**

菲利普·赫费尔　**PHILIPP HOEVEL**

罗伯塔·西纳特拉　**ROBERTA SINATRA**

1.1 哥尼斯堡的桥

很少有哪个研究领域的诞生可以精确地追溯到历史上的某个时刻和地方。而图论——网络科学的数学基础，做到了这一点。图论起源于1735年的哥尼斯堡（Königsberg）。作为东普鲁士的首府，哥尼斯堡在那个时代是一座繁荣的商业城市。由络绎不绝的商船支撑着的贸易，为哥尼斯堡带来了丰厚的经济收入。经费充足的市政府在环绕市区的普雷格尔河（Pregel）上建造了7座桥。普雷格尔河的两条支流将美丽的奈佛夫岛（Kneiphof）和大陆隔开，奈佛夫岛和大陆通过5座桥相连。另外的两座桥分别修建在普雷格尔河的两条支流上（图**1-1a**）。如此独特的城市布局给哥尼斯堡人带来了一个时代难题：能否不重复地走过每座桥？人们做了很多次尝试，却始终没有找到这样的路径。这个问题一直无人能解，直到1735年，出生于瑞士的数学家莱昂哈德·欧拉（Leonhard Euler）才对此做出了严格的数学证明：不存在这样的路径[6]，[7]。

欧拉使用字母A、B、C、D表示被河流隔开的4块陆地（图**1-1b**）。接下来，对于连接两块陆地的每座桥，使用一个链接来表示。最终，他绘制出一幅图（图**1-1c**），图中的节点是陆地，链接对应着桥。然后欧拉做了一个简单的观察：如果存在一条经过每座桥各一次的路径，那么拥有奇数个链接的节点只能是该路径的起点或终点。实际上，当你到达一个拥有奇数个链接的节点时，你可能会发现，已经没有未走过的链接让你离开那里了。

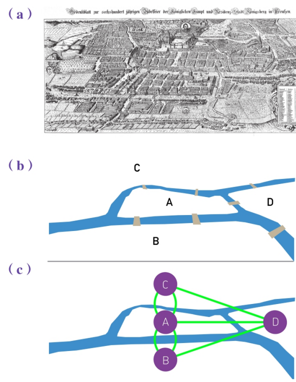


图1-1　哥尼斯堡的桥

（a）欧拉那个时代的哥尼斯堡（如今是俄罗斯的加里宁格勒市B）地图。

（b）哥尼斯堡4块陆地和7座桥的示意图。

（c）　欧拉构造了一个有4个节点（A、B、C、D）和7个链接的图，每个节点对应着一块陆地，7个链接则对应着7座桥。欧拉证明了，在哥尼斯堡不存在经过每座桥各一次的路径。哥尼斯堡人也因此放弃了他们注定没有结果的寻找，于1875年在B和C之间建造了一座新桥。这座桥使节点B和C的链接数增加到了4个。现在，只有两个节点拥有奇数个链接了。如此一来，我们就可以找到想要的路径。你找到了吗？

经过每座桥各一次的路径只能有一个起点和一个终点。因此，如果一幅图中拥有奇数个链接的节点超过了2个，那么图中就不存在这样的路径。哥尼斯堡七桥所对应的图中，拥有奇数个链接的节点有4个，分别是A、B、C和D，因此不存在满足要求的路径。

欧拉的证明第一次使用了图来求解数学问题。这一证明给我们带来了两点启示：首先，有些问题在使用图进行表示时会变得简单起来，且更容易求解；其次，路径的存在性是图的一种性质，不取决于我们是否能够找出这样的路径。实际上，在哥尼斯堡图结构已知的情况下，无论我们多么聪明，都无法找到想要的路径。换句话说，网络的某些性质是编码在其结构中的，网络性质和其结构密切相关。

要理解网络如何能够影响其所描述系统的性质，我们需要熟悉和掌握这个从欧拉的证明中成长起来的数学分支——图论。这一章介绍度、度分布、路径、距离等网络基本概念，我们要学习如何将网络表

示成图，了解加权图、有向图和二分图。本章还介绍了本书将会使用的图论形式化方式和语言。



在线资源**1.1**　哥尼斯堡桥

扫描二维码下载“湛庐阅读”App，回复书名，观看介绍哥尼斯堡七桥问题和欧拉解法的视频。

1.2
网络和图

要理解复杂系统，首先需要知道其各个组成部分之间彼此交互的方式。换句话说，我们需要一幅刻画各组成部分之间连接关系的地图。网络记录了系统各组成部分（节点或顶点）及它们之间的连接关系（被称为链接或边），见边栏**1.1**　。网络作为复杂系统的一种表示方式，为研究性质、外观、范畴各异的复杂系统提供了公共语言。实际上，如图**1-2**　所示，三个截然不同的系统却有着完全相同的网络表示。

边栏**1.1**

网络，还是图？

在科技文献中，关于网络和图的一些词语经常被人们不加区分地交替使用（本书对这些词的使用也是如此）。例如：网络科学和图论，网络和图，节点和顶点、链接和边。

不过，网络和图之间还是有一些细微差别的。“网络、节点、链接”这个组合通常用于表示真实的系统：万维网是网页之间的网络，社会是个体间通过家庭、朋友、职业纽带等连接起来的网络；代谢网络是细胞内发生的所有化学反应的总和。相对而言，“图、顶点、边”则经常在讨论这些网络的数学表示时使用：万维网图、社会关系图（因Facebook而流行起来的一个词）或代谢关系图。然而，人们很少会对其进行区分，因此这些术语经常被作为同义词使用。

图**1-2** 介绍了两个基本的网络参数：

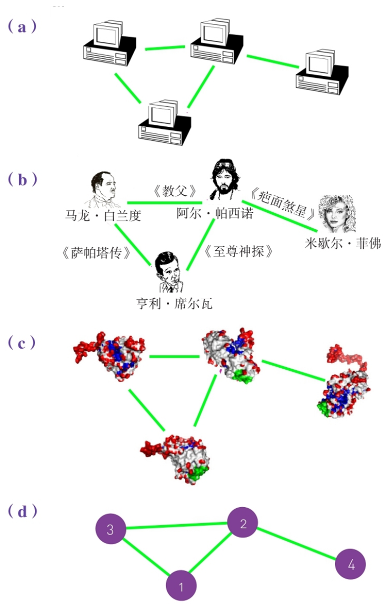


图**1-2**　不同的网络，同样的图

图中展示了三个不同网络的一部分。（a）互联网：路由器彼此互联；（b）好莱坞演员网络：出演过同一部影片的两个演员之间相互连接；（c）蛋白质相互作用网络：细胞中可以拼接在一起的两个蛋白质之间彼此连接。虽然这些网络的节点和链接

性质不同，但它们可以用同样的图来表示，由4个节点和4个链接构成，如（d）所示。

节点数N ，表示系统中组成部分的个数。我们通常将N 称为网络大小。为便于区分，网络中的各个节点记为$i =1,2, \ldots, N$ 。

链接数L ，表示节点间交互关系的总数。一般而言，我们很少对链接直接进行标记，而是通过其连接的两个节点来标记。例如，（2, 4）表示连接节点2和节点4的链接。

图**1-2** 中的网络，节点数和链接数均为4。

网络中的链接可以是无向的，也可以是有向的。有些系统中的链接是有向的。例如，万维网中的链接由一个网页指向另一个网页；手机通话网络中的链接由呼叫者指向被呼叫者。有些系统中的链接是无向的。例如，恋爱关系网络中，恋爱双方之间的关系是无向的；电网中的传输线是无向的，电流可以沿着两个方向流动。

如果一个网络中所有的链接都是有向的，我们称之为有向网络或有向图。同样，如果其所有链接都是无向的，我们称之为无向网络。还有一些网络同时包含有向的链接和无向的链接。例如，在代谢关系网络中，有些代谢关系是可逆的——对应无向的链接，有些代谢关系是不可逆的——对应有向的链接。

在使用网络来表示一个系统时，我们所做的选择将决定我们使用网络科学成功求解特定问题的能力。例如，定义两个个体之间链接的方式决定了我们所能研究的问题：

（**1**） 把工作中经常交互的个体连接起来所得到的职业关系网络，对于公司或组织的成功具有重要作用，受到了组织行为学研究的大量关注。

（**2**） 将好友连接在一起而得到的好友关系网络，在观点、产品和习惯的传播中扮演着重要角色，引起了社会学、营销学和健康科学的研究兴趣。

（**3**） 通过将具有亲密关系的个体连接起来而得到的性关系网络，对于研究艾滋病等性传染疾病的传播具有重要意义，受到了传染病研究的密切关注。

（**4**） 根据邮件和通话记录构建的熟识关系网络，是上述职业关系、朋友关系、亲密关系的混合，对于研究交流和营销具有重要意义。

上述4个网络虽然有很多重合的链接（有些同事之间也是朋友关系，甚至有亲密关系），它们的用途和构造目的却大不相同。

我们也可以构造出一些从图论角度看是合法的、却没有什么实际用途的网络。例如，将同姓的人连接起来可以得到一个从定义上看没什么问题的网络（约翰和约翰相连，而玛丽和玛丽相连），我们可以使用网络科学工具进对其性质进行分析。然而，其实际用途却是值得怀疑的。因此，将网络科学用于研究复杂系统时，需要仔细考虑节点和链接的选择，以确保构造出的网络对于所研究的问题是有意义的。

本书将使用10个网络作为例子来阐述网络科学工具。表**1-1** 列出了这些参照网络，涉及社会系统——手机通话网络和电子邮件网络、合作网络（科学合作关系网络、好莱坞演员关系网络）、信息系统（万维网）、技术和基础设施系统（互联网、电网）、生物系统（蛋白质相互作用网络、代谢关系网络）和引文网络（科学引文网络）。这些网络的大小差异很大——小到只有1 039个节点的大肠杆菌代谢关系网络，大到包含几十万个节点的引文网络。这些网络涵盖了网络科学的多个活跃领域，是研究人员经常用来展示网络关键性质的例子。如表**1-1** 所示，这些网络中有的是有向的，有的是无向的。在接下来的章节中，我们将详细探讨每个网络的特性，这些网络在我们理解复杂网络的旅程中将扮演小白鼠的角色。

表**1-1** 经典网络数据

网络	节点	链接	有向或无向	N	L	$\langle k \rangle$
互联网	路由器	网络连接	无向	192 244	609 066	6.34
万维网	网页	链接	有向	325 729	1 497 134	4.60
电网	发电厂、变压器	电缆	无向	4 941	6 594	2.67
手机通话网络	用户	通话	有向	36 595	91 826	2.51
电子邮件网络	电子邮箱地址	电子邮件	有向	57 194	103 731	1.81
科学合作网络	科学家	共同研究	无向	23 133	93 439	8.08
演员网络	演员	共同表演	无向	702 388	29 397 908	83.71
引文网络	论文	引用	有向	449 673	4 689 479	10.43
大肠杆菌代谢网络	代谢分子	化学反应	有向	1 039	5 802	5.58
蛋白质相互作用网络	蛋白质	结合交互	无向	2 018	2 930	2.90

本书用于阐述网络科学工具的10个经典网络。表中列出了每个网络中节点和链接的性质、链接的方向性、节点个数、链接个数和平均度。对于有向网络，平均度是平均入度和平均出度的均值（见公式**1.5**）。

1.3 度、平均度和度分布

度是节点的一个关键属性，表示该节点和其他节点之间的链接个数。在手机通话网络中，度表示一个人通过手机联系过的人数；在引文网络中，度表示一篇论文获得的引用数。

度

我们使用k_i表示网络中第i个节点的度。例如，对于图**1-2**所示的无向网络，我们有$k_1=2$，$k_2=3$，$k_3=2$，$k_4=1$。在无向网络中，总链接数L可以使用节点度之和来表示：

$$L=\frac{1}{2}\sum_{i=1}^{N}k_i \tag{1.1}$$

在计算节点度之和时，每个链接被计算了两次，因此公式**1.1**中有一个系数1/2。例如，在图**1-2**中，连接节点2和节点4的链接，在计算节点2的度时被计算了一次，在计算节点4的度时又被计算了一次。

平均度

平均度是网络的一个重要属性（边栏**1.2**）。在无向网络中，平均度被定义为：

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \qquad (1.2)$$

边栏**1.2**

统计知识简要回顾

对于N 个样本$x_1, \ldots, x_N$，4个重要的统计量：

均值：

$$\langle x \rangle = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

n 阶矩：

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \cdots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

标准差：

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

样本分布：

$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

这里p_x 满足归一化约束

$$\sum_i p_x = 1\left(\int p_x dx = 1\right)$$

有向网络中，需要区分入度和出度。入度k_i^{in} 表示指向节点i 的链

接个数，而出度k_i^{out} 表示节点i 指向其他节点的链接个数。从而，节点i 的度为：

$$k_i = k_i^{\text{in}} + k_i^{\text{out}} \tag{1.3}$$

例如，在万维网中，网页的出度k^{out} 表示该网页指向的网页数目，其入度k^{in} 表示指向它的网页数目。有向网络中的总链接数为：

$$L = \sum_{i=1}^{N} k_i^{\text{in}} = \sum_{i=1}^{N} k_i^{\text{out}} \tag{1.4}$$

这里不像公式**1.1** 那样有一个系数1/2：公式**1.4** 的两个求和项将出度和入度分开进行计算。有向网络的平均度为：

$$\langle k^{\text{in}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k^{\text{in}} = \langle k^{\text{out}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k^{\text{out}} = \frac{L}{N} \tag{1.5}$$

度分布

度分布p_k 表示"网络中随机选出的一个节点其度为k "的概率。由于p_k 是一个概率，则其必须满足归一化约束，即：

$$\sum_{k=0}^{\infty} p_k = 1 \tag{1.6}$$

对于有N 个节点的网络而言，其度分布可以表示成归一化的直方图（图**1-3** ）：

$$p_k = \frac{N_k}{N} \tag{1.7}$$

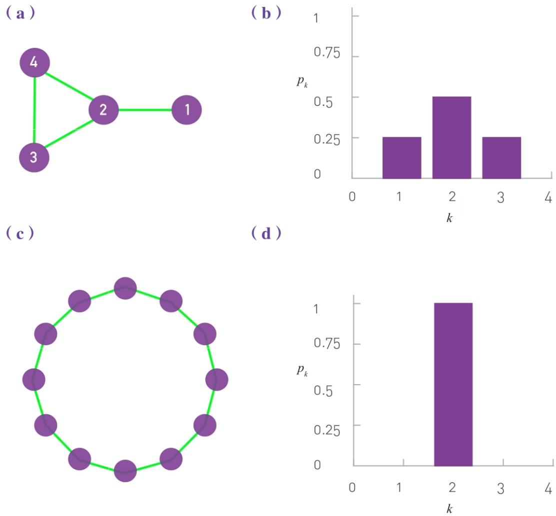


图1-3　度分布

网络的度分布可以由公式**1.7** 计算得到。

（**a**） 该网络大小为$N=4$，其度分布如图b所示。

（**b**） 4个节点中有一个度为1的节点（$k_1=1$），因此$p_1=1/4$；有两个度为2的节点（$k_3=k_4=2$），因此$p_2=1/2$；有一个度为3的节点（$k_2=3$），因此$p_3=1/4$。没有度大于3的节点，因此，对于任意$k>3$，我们有$p_k=0$。

（**c**） 对于一维格子网络而言，所有节点的度都是2。

（**d**） 图c所示网络的度分布是克罗内克（Kronecker）的德尔塔函数，即$p_k=\delta(k-2)$。

这里的N_k 是指度为k 的节点个数。因此，度为k 的节点个数可以根据度分布计算出来，即$N_k=Np_k$ 。

度分布被认为在网络科学中发挥了核心作用，特别是在无标度网络被发现之后[8]。其中一个原因是，大多数网络性质可以通过度分布计算得到。例如，网络平均度可以写成：

$$\langle k\rangle=\sum_{k=0}^{\infty}kp_k \qquad (1.8)$$

另一个原因是，度分布的具体函数形式决定着很多网络现象，例如网络健壮性和病毒传播。接下来，让我们看一看真实网络中有着怎

样的度分布（图**1-4** ）。

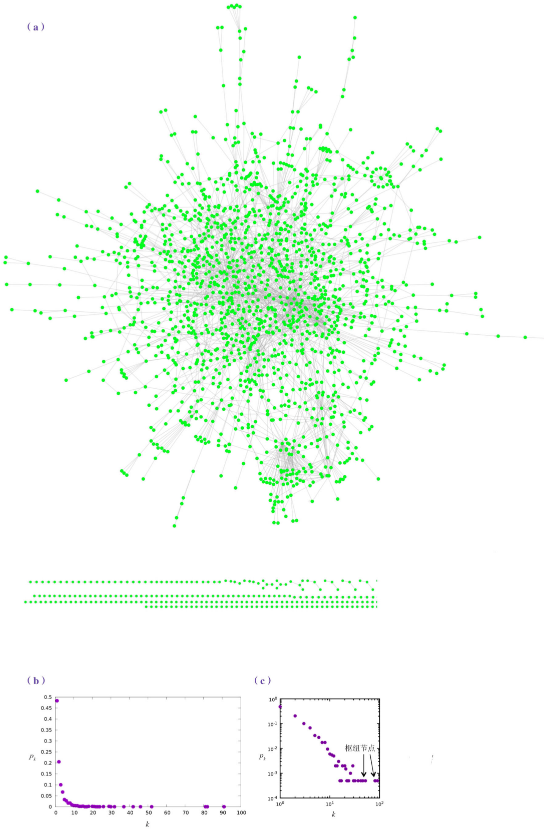


图1-4　一个真实网络的度分布

真实网络中，节点的度差异很大。

（a）　酵母的蛋白质相互作用网络（表**1-1**　）。每个节点对应一个酵母蛋白质，链接表示实验发现的蛋白质间的相互作用。注意，底部所示的几个蛋白质有自环，它们的度为2。

（b）　图a所示的蛋白质相互作用网络的度分布。节点的度分布在0（对应孤立节点）和92（对应度最大的那个节点，此类节点通常被称为枢纽节点）之间。度不同的节点，其个数也差异很大：度为1的节点有几乎一半（$p_1 = 0.48$），而只有一个节点的度为92（$p_{92} = 1/N = 0.000\,5$）。

（c）　度分布通常画在双对数坐标下。当然，除了采用双对数坐标之外，也可以直接绘制$\ln p_k$ 关于$\ln k$ 的函数。采用双对数坐标的好处会在第**3**章 进行讨论。

1.4 邻接矩阵

完全描述一个网络需要记录下它的所有链接。实现这一目标的最简单方式是使用列表列出所有的链接。例如，图**1-2** 中的网络可以通过列出其全部4个链接来进行唯一刻画：{(1,2),(1,3),(2,3),(2,4)}。从数学角度考虑，我们经常使用邻接矩阵来表示网络。一个拥有N 个节点的有向网络，其邻接矩阵有N 行N 列，邻接矩阵的元素定义如下：如果存在一个链接从节点j 指向节点i ，则$A_{ij}=1$；如果节点i 和节点j 彼此不相连，则$A_{ij}=0$。

对于无向网络，每个链接对应邻接矩阵中的两个元素。例如，链接（1,2）对应着$A_{12}=1$和$A_{21}=1$。因此，无向网络的邻接矩阵是对称的，即$A_{ij}=A_{ji}$（图**1-5b** ）。

节点i 的度k_i 可以从邻接矩阵的元素中直接得到。对于无向网络，一个节点的度是邻接矩阵中相应行或列中所有元素之和，即：

$$k_i=\sum_{j=1}^{N}A_{ij}=\sum_{j=1}^{N}A_{ji} \tag{1.9}$$

对于有向网络，邻接矩阵中一行元素之和及一列元素之和分别对应着节点的入度和出度，即：

$$k_i^{\text{in}}=\sum_{j=1}^{N}A_{ij},\ k_i^{\text{out}}=\sum_{j=1}^{N}A_{ji} \tag{1.10}$$

对于无向网络，节点的入度和节点的出度相等，我们有：

$$2L=\sum_{i=1}^{N}k_i^{\text{in}}=\sum_{i=1}^{N}k_i^{\text{out}}=\sum_{ij}^{N}A_{ij} \tag{1.11}$$

无向网络的邻接矩阵中非零元素个数为$2L$ ，即链接个数的2倍。实际上，连接节点i 和节点j 的一个无向链接对应着邻接矩阵中的两项：表示从节点j 指向节点i 的$A_{ij}=1$和表示从节点i 指向节点j 的$A_{ji}=1$（图**1-5b** ）。

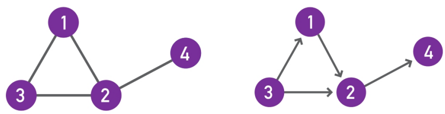
（a）邻接矩阵

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \end{pmatrix}$$

（a）邻接矩阵

$$A_{ij}=\begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$

（b）无向网络　　　　（c）有向网络



$$A_{ij}=\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad A_{ij}=\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$k_2=\sum_{j=1}^{4}A_{2j}=\sum_{i=1}^{4}A_{i2}=3 \qquad k_2^{\text{in}}=\sum_{j=1}^{4}A_{2j}=2，\ k_2^{\text{out}}=\sum_{i=1}^{4}A_{i2}=1$$

$$A_{ij}=A_{ji} \quad A_{ii}=0 \qquad A_{ij}\neq A_{ji} \quad A_{ii}=0$$

$$L=\frac{1}{2}\sum_{i,j=1}^{N}A_{ij} \qquad L=\sum_{i,j=1}^{N}A_{ij}$$

$$\langle k\rangle=\frac{2L}{N} \qquad \langle k^{\text{in}}\rangle=\langle k^{\text{out}}\rangle=\frac{L}{N}$$

图1-5　邻接矩阵

（**a**）邻接矩阵元素的标记方式。

（**b**）无向网络的邻接矩阵。如图所示，一个节点（例如节点2）的度可以表示为邻接矩阵中相应行或列的元素之和。图中还展示了一些可以从邻接矩阵中得到的网络性质，例如链接总数L和平均度$\langle k\rangle$。

（**c**）有向网络的邻接矩阵。

1.5
真实网络是稀疏的

现实世界中，不同网络的节点数（N）和链接数（L）差异很大。例如，秀丽隐杆线虫的神经元网络——到目前为止唯一完整得到的生物体神经系统，拥有$N=302$个神经元（节点）。与之对应，人

类大脑据估计大约有1000亿个神经元（$N \approx 10^{11}$）。人类细胞的基因网络有大约2万个基因；社会网络中约有70亿个个体（$N \approx 7 \times 10^9$）；万维网中据估计有超过1万亿个网页（$N > 10^{12}$）。

网络大小上的巨大差异可以从表**1-1** 中明显看出。表**1-1** 列出了一些网络的节点数和链接数。有些网络是其所描述系统的完整连接关系，例如演员关系网络和秀丽隐杆线虫的神经元网络；而其他网络则只是整个网络的一部分，例如万维网和手机通话网络。

从表**1-1** 中可以看出，网络的链接数差异也很大。从理论上讲，对于拥有N 个节点的无向网络，其链接数最小为$L=0$，最大为L_{max}。其中，

$$L_{max} = \binom{N}{2} = \frac{N(N-1)}{2} \qquad (1.12)$$

表示大小为N 的完全图包含的链接总数（图**1-6** ）。在完全图中，任意两个节点之间都有链接相连。

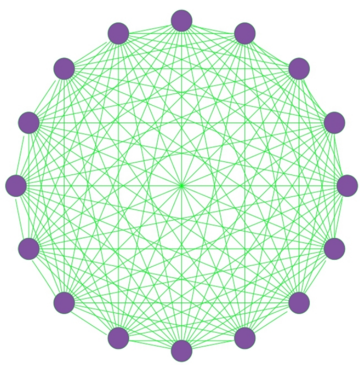


图**1-6** 完全图

包含$N=16$个节点的完全图，根据公式**1.12** ，其链接数为$L_{max}=120$。完全图的邻接矩阵中，对角线元素为0（$A_{ii}=0$），其他元素都为1：$A_{ij}=1(i \neq j)$。完全图的平均度为$\langle k \rangle = N-1$。完全图通常被称为团（clique），第**8**章 探讨社区识别问题时会经常使用这个词。

真实网络中，L 远小于L_{max}，说明大部分真实网络是稀疏的。当

$L \ll L_{max}$ 时，我们说该网络是稀疏的。例如，表**1-1** 中的万维网包含大约150万个链接。不过，如果该万维网是完全图，根据公式**1.12** ，它应该有$L_{max} \approx 5 \times 10^{10}$ 个链接。也就是说，万维网实际包含的链接数占最大可能链接数的比例仅为3×10^{-5} 。实际上，表**1-1** 中的所有网络都是稀疏的：实际链接数相对于最大可能链接数（具有同样节点数的完全图所包含的链接数）只占很小一部分比例。

真实网络的稀疏性意味着其邻接矩阵也是稀疏的。实际上，完全图的邻接矩阵中，所有非对角线元素都为1：$A_{ij} = 1(i \neq j)$。真实网络的邻接矩阵中，只有很小一部分元素非零。图**1-7** 展示了表**1-1** 列出的蛋白质相互作用网络（图**1-4a** ）的邻接矩阵。可以看出，该矩阵几乎是空的。

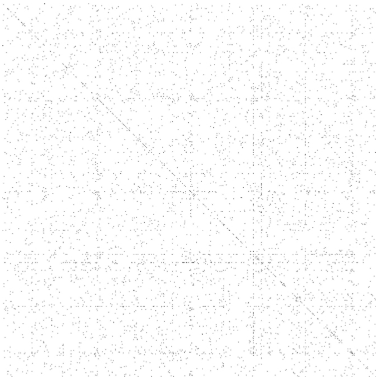


图**1-7** 稀疏邻接矩阵

酵母蛋白质相互作用网络的邻接矩阵。该网络包含2018个节点，每个节点表示一个酵母蛋白质（表**1-1** ）。图中的每个点对应着邻接矩阵中的一个非零元素，表示存在一个蛋白质相互作用。对角线上没有点出现，因为所有对角线元素为0。只有很少的地方有点出现，体现了蛋白质相互作用网络的稀疏性。

稀疏性对于我们研究和存储真实网络有着重要意义。例如，在计算机中存储一个大网络时，只存储链接列表（邻接矩阵中的非零元素）相对于存储整个邻接矩阵会好很多，因为邻接矩阵中相当大比例的元素都是0。因此，使用矩阵表示网络时，零元素会浪费大量的内存空间（图**1-7** ）。

1.6 加权网络

到目前为止，我们探讨的网络中，所有链接的权重都一样（$A_{ij}=1$）。然而，在许多应用中，我们需要研究加权网络。加权网络中的每个链接（i，j）都有自己的权重w_{ij}。在手机通话网络中，权重可以表示两个人之间手机通话的时长（例如分钟数）；在电网中，权重可以表示传输线路上流过的电流量。

对于加权网络而言，邻接矩阵的元素表示链接的权重，即

$$A_{ij}=w_{ij} \tag{1.13}$$

令科学界感兴趣的大多数网络都是加权网络，但我们并不总能找到其合适的权重。因此，我们经常使用无权网络来近似加权网络。本书中，我们主要关注无权网络，但会在合适的时候讨论权重对网络性质的影响（边栏**1.3**）。

边栏**1.3**

梅特卡夫定律：网络的价值

梅特卡夫定律指出，网络的价值与网络节点个数的平方（N^2）成正比。该定律是罗伯特·梅特卡夫（Robert M. Metcalfe）在1980年左右阐述通信设备网络价值时提出的[9]。梅特卡夫定律背后的思想是，网络用户数越多，网络就越有价值。例如，你的朋友中使用电子邮件的人越多，电子邮件这项服务对你而言就越有价值。

20世纪90年代后期互联网泡沫期间，梅特卡夫定律经常被用于对互联网公司进行定量估值。梅特卡夫定律表明，一项互联网服务的价值和其能创造的链接数成正比，也就是和其用户数的平方成正比。相比而言，其成本和用户数（N）却只是线性相关。因此，对互联网服务而言，只要能够吸引足够多的用户，就一定能够赢利

——随着N的增加，N^2总会超过N（图1-8）。梅特卡夫定律营造了“有服务就不愁用户”的心态，支撑着“增长必将带来利润”的信念。

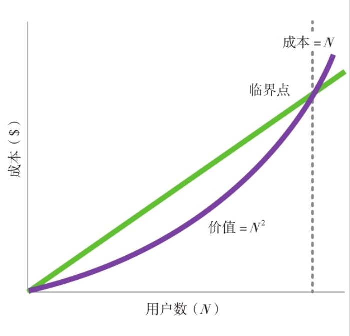


图1-8　梅特卡夫定律

根据梅特卡夫定律，基于网络的服务或技术 - 成本随着节点（用户或设备）数目线性增长。相比之下，根据公式**1.12**，由链接数目L_{max}决定的收益或利润，其增速与N^2成正比。因此，一旦用户数或设备数超过某个临界值，这项新服务或新技术就会赢利。

梅特卡夫定律是基于公式**1.12**的：对于拥有N个用户的通信网络，如果其每个链接的价值都是一样的，则网络总价值正比于$N(N-1)/2$，大致为N^2。如果一个网络有$N=10$个用户，那么这些用户之间最多可有$L_{max}=45$个不同的链接。如果网络大小增加一倍（$N=20$），链接数会增加不止一倍，而是差不多4倍（$L_{max}=190$）。这一现象在经济学中被称为网络外部性（network externality）。

有两个问题限制了梅特卡夫定律的有效性：

（**1**）　大多数网络是稀疏的，这意味着只有很小一部分链接会出现。因此，网络价值并非像N^2那样增大，而是与网络大小N呈线性关系。

（**2**）　链接是有权重的，并不是所有链接的价值都一样。有些链接经常被使用，而绝大多数链接很少被使用。

1.7 二分网络

二分网络（或二分图）是这样一种网络：其节点可以分成两个不相交的节点集合U和V，使得每个链接都连接着一个U中的节点和一个V中的节点。换句话说，如果我们把集合U中的节点涂成绿色，把集合V中的节点涂成紫色，则每个链接连接着颜色不同的两个节点（图**1-9**）。

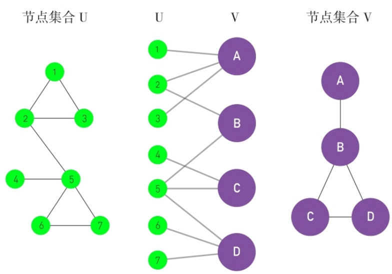


图**1-9**　二分网络

一个二分网络包含两个节点集合U和V。集合U中的节点只和集合V中的节点直接相连。因此，集合U的节点之间不存在直接链接，集合V的节点之间同样也不存在直接链接。如图所示，任意二分网络都可以生成两个映射网络。其中一个映射网络映射到节点集合U上，如果U中的两个节点在二分网络中有链接指向V中同样的节点，则这两个节点在该映射网络中彼此相连。另一个映射网络映射到节点集合V上，如果V中的两个节点在二分网络中有链接指向U中同样的节点，则这两个节点在该映射网络中彼此相连。

对于每个二分网络，我们可以生成两个映射网络。第一个映射网络中，节点集合是U，如果U中的两个节点都有链接指向V中的同一个节点，则它们在该映射网络中有链接相连。与之类似，第二个映射网络对应着节点集合V上的一个网络（图**1-9**）。

网络科学研究中，我们会碰到很多二分网络。一个著名的例子是好莱坞电影演员网络，其中一个节点集合对应着电影（记为U），另一个节点集合对应着演员（记为V）。如果某个演员出演了某部电影，则二者之间有一个链接。该二分网络的一个映射网络是演员网络

——出演过同一部电影的两个演员之间彼此相连（表**1-1**　中列出的演员网络）。另一个映射网络是电影网络——至少有一个演员相同的两部电影彼此相连。

医药领域也有一个明显的二分网络：人类疾病网络。基因和疾病构成该二分网络的节点，链接刻画基因和疾病之间的因果关系（图**1-10**　）。最后，我们还可以定义多分网络，例如由食谱、原料、成分构成的三分网络（图**1-11**　）。

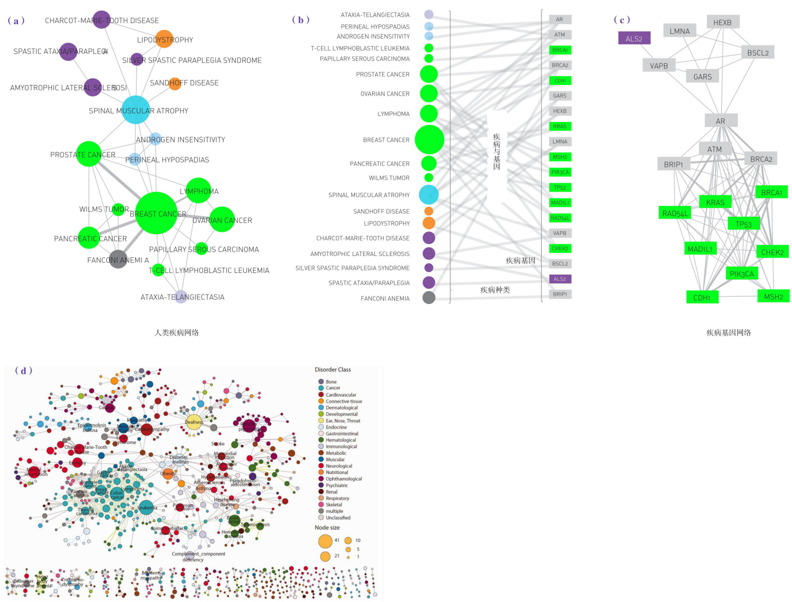


图**1-10**　人类疾病网络

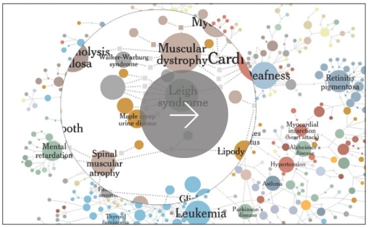
（**a**）　人类疾病网络的一个映射网络是基因网络：节点是基因，如果两个基因和同一种疾病相关，则它们在该映射网络中彼此相连。

（**b**）　人类疾病网络是一个二分网络，包括疾病（记为U）和基因（记为V）两类节点。如果某个基因的突变导致了某种疾病，则它们对应的两个节点之间彼此相连[4]。

（**c**）　人类疾病网络的第二个映射是疾病网络：节点是疾病，如果两种疾病和同样的基因相关——由同样的基因突变造成，则它们在该映射网络中彼此相连。图a~c展示的是人类疾病网络的一个子集——和癌症相关的子集。

（**d**）完整的人类疾病网络包含1 283种疾病和1 777个疾病基因[1]。详细的人类疾

病网络参见在线资源**1.2**。



在线资源**1.2**　人类疾病网络

在这里可下载人类疾病网络的高分辨率版本 [1] ，或者使用《纽约时报》构建的在线接口浏览该网络。

扫描二维码下载“湛庐阅读”App，回复书名，下载人类疾病网络的高分辨率版本。

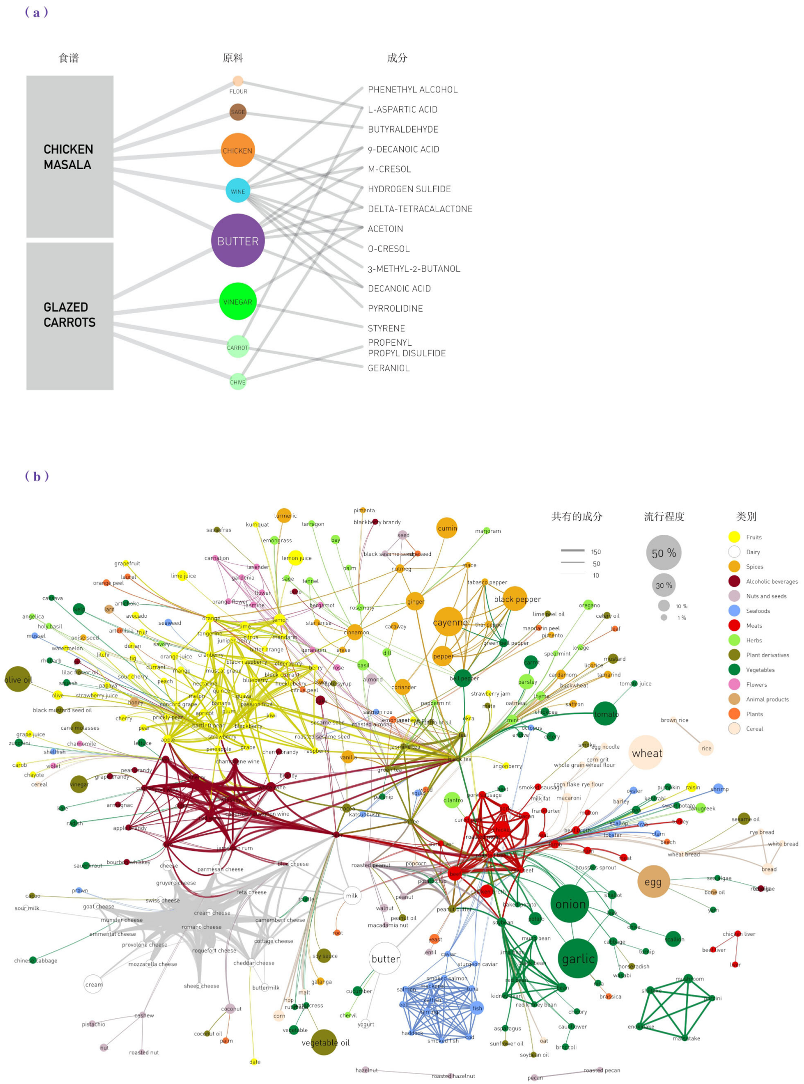


图1-11　三分网络

（a）　“食谱 - 原料 - 成分”三分网络的构造方式：第一个节点集合是食谱，例如马沙拉鸡（Chicken　Marsala）；第二个节点集合对应食谱的原料（例如，做马沙拉鸡所需的面粉、鼠尾草、鸡、白酒、黄油等原料）；第三个节点集合表示成分——体现原料味道的化学成分。

（b）原料网络是图a所示三分网络的一个映射网络。每个节点表示一种原料，节点颜色表示食物的类别，节点大小表示原料的流行程度。如果两种原料有许多

相同的成分，则它们对应的节点彼此相连。链接的粗细表示两种原料共同拥有的成分数目多少 [11] 。

1.8
路径和距离

物理距离在物理系统各组成部分间的交互中起着关键作用。例如，晶体中两个原子间的距离、宇宙中两个星系间的距离决定着它们之间作用力的大小。

网络距离是一个比较难理解的概念。什么是两个网页之间的距离？互不相识的两个人之间的距离是多少？物理距离在这里不再适用：两个网页可能放置在位于地球上位置不同、方向相反的两台电脑上，但它们之间有一个链接使彼此相连。同时，生活在同一栋楼上的两个人可能互不认识。

网络中，物理距离被路径长度取代了。路径是指沿着网络中的链接行走经过的线路。路径的长度表示其包含的链接个数（图**1-12a**）。注意，有些情况下会要求一条路径经过每个节点不能超过一次。

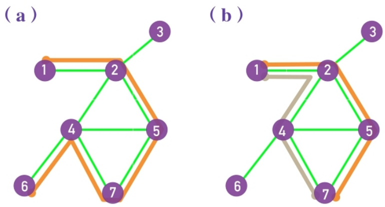


图**1-12**　路径

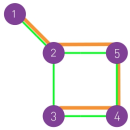
（**a**） 节点i_0 和节点i_n 之间的一条路径是包含n 个链接的一个有序列表$P = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$。路径的长度为$n$ 。图a中的橙色路径对应的线路为1→2→5→7→4→6，路径长度为$n = 5$。

（**b**） 节点1和节点7之间的最短路径对应着连接节点1和节点7的、包含最少链接数的路径。如图中的橙色路径和灰色路径所示，存在多条长度相同的路径。网络直径是网络中所有节点对之间的最大距离，这里的网络直径为$d_{max} = 3$。

在网络科学中，路径扮演着核心角色。接下来，我们将探讨一些

关于路径的重要性质（更多介绍参见图1-13 ）。

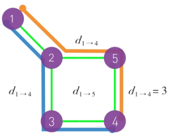
（a）



路径

一个节点序列，序列中每个节点有链接指向其后面紧邻的节点。每个路径包含 $n+1$ 个节点和 n 个链接。路径的长度是其包含的链接数，多次出现的链接重复计数。例如，橙色路径 $1\to2\to5\to4\to3$ 的路径长度为 4。

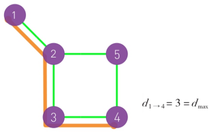
（b）



最短路径

两个节点间长度最短的路径。我们将两个节点之间最短路径的长度称为测地距离 d。注意，最短路径不一定是唯一的：节点 1 和节点 4 之间有两条最短路径，分别是 $1\to2\to3\to4$（蓝色）和 $1\to2\to5\to4$（橙色），它们的路径长度都为 $d_{1,4}=3$。

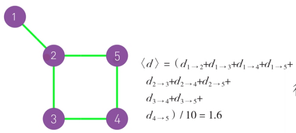
（c）



直径（d_{max}）

最短路径的最大长度，也是距离最远的节点对之间的距离。如图所示，节点 1 和节点 4 之间的距离最远，因此网络直径为 $d_{max}=3$。

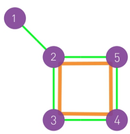
（d）



平均路径长度（$\langle d\rangle$）

所有节点对之间最短路径的平均长度。如图所示，该网络的平均路径长度为 $\langle d\rangle=1.6$，计算过程列在了图的旁边。

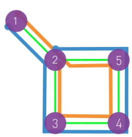
（e）



环

起始节点和终止节点相同的路径。如图所示，橙色路径是一个环。

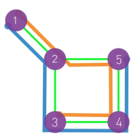
（f）



欧拉路径

经过每条链接恰好一次的路径。图中给出了两条欧拉路径，分别为橙色和蓝色的。

（g）



汉密尔顿路径

经过每个节点恰好一次的路径。图中给出了两条汉密尔顿路径，分别为橙色和蓝色的。

图1-13　路径学

最短路径

节点i 和节点j 之间的最短路径是指包含链接数最少的路径（图**1-12b**）。最短路径的长度通常被称为节点i 和节点j 之间的距离，记为d_{ij} 或简略记为d 。同一对节点之间可能有多条长度相同的最短路径（图**1-12b**）。注意，最短路径不会包含环或自环。

在无向网络中，我们有$d_{ij}=d_{ji}$ ，即节点i 和节点j 之间的距离等于节点j 和节点i 之间的距离。在有向网络中，通常情况下$d_{ij}\neq d_{ji}$ 。另外，有向网络中，存在节点i 到节点j 的路径并不意味着存在节点j 到节点i 的路径。

在真实网络中，我们通常需要确定两个节点之间的距离。对于小网络而言，譬如图**1-12** 所示的网络，这是一件容易的事情。不过，对于有数百万个节点的网络而言，找出两个节点之间的最短路径就相当费时了。最短路径的长度和最短路径的个数可以从邻接矩阵得到（边栏**1.4**）。实际上，我们通常采用边栏**1.5** 中讨论的广度优先搜索算法来实现。

边栏1.4

两个节点间的最短路径数

节点i 和节点j 之间的距离d_{ij} 和最短路径数N_{ij} 可以从邻接矩阵A_{ij} 直接计算得到。

$d_{ij}=1$：如果节点i 和节点j 之间有直接链接（$A_{ij}=1$），则$d_{ij}=1$。

$d_{ij}=2$：如果节点i 和节点j 之间存在长度为2的路径，假如该路径经过的中间节点为k ，则有$A_{ik}A_{kj}=1$；否则$A_{ik}A_{kj}=0$。因此，节点i 和节点j 之间长度为2的路径数为

$$N_{ij}^{(2)}=\sum_{k=1}^{N}A_{ik}A_{kj}=\left[A^2\right]_{ij}$$

这里$[\ldots]_{ij}$ 表示矩阵第i 行第j 列的元素。

$d_{ij}=d$ ：如果节点i 和节点j 之间存在长度为d 的路径，则该路

径包含的链接所对应的邻接矩阵元素相乘为1，即$A_{ik}\ldots A_{lj}=1$；否则$A_{ik}\ldots A_{lj}=0$。因此，节点i 和节点j 之间长度为d 的路径数为：

$$N_{ij}^{(d)}=\left[A^d\right]_{ij}$$

上述公式对于有向网络和无向网络都适用。节点i 和节点j 之间的距离是满足$N_{ij}^{(d)}>0$ 的最小的d （注意，这里的路径不能包含环或自环）。虽然该方法很简洁，但对于大的网络，更高效的方法是边栏**1.5** 描述的广度优先搜索算法。

网络直径

网络直径，记为$d_{\max}$ ，是指网络中所有最短路径的最大长度。换句话说，网络直径是指网络中所有节点对之间的最大距离。图**1-13** 所示的网络，其直径为$d_{\max}=3$。对于较大的网络，网络直径可以使用边栏**1.5** 中讨论的广度优先搜索算法计算得到。

平均路径长度

平均路径长度，表示为$\langle d\rangle$ ，是指网络中所有节点对之间的平均距离。对于包含N 个节点的有向网络，我们有

$$\langle d\rangle=\frac{1}{N(N-1)}\sum_{\substack{i,j=1,\cdots,N\\ i\neq j}} d_{ij}. \tag{1.14}$$

注意，公式**1.14** 只测量出现在同一个网络连通分支中的节点对（**1.9**节 ）。我们可以使用广度优先搜索算法来计算一个大网络的平均路径长度。为此，首先使用边栏**1.5** 中的算法计算第一个节点和其他所有节点的距离，再计算第二个节点和其他所有节点（对于无向网络，第一个节点除外）的距离，然后再对其他所有节点重复该过程。

边栏**1.5**

广度优先搜索算法

广度优先搜索（BFS）算法是网络科学中经常使用的一个算法。向池塘里扔一块石头，然后观察扩散开的涟漪，这就是BFS算法的思想。BFS算法从一个源节点出发，标记其邻居（第一个涟漪），然后是邻居的邻居（第二个涟漪），直到到达目标节点。到达目标节点之前需要的“涟漪”数就是源节点和目标节点间的距离。

寻找节点i 和节点j 之间最短路径的步骤如下（图**1-14** ）：

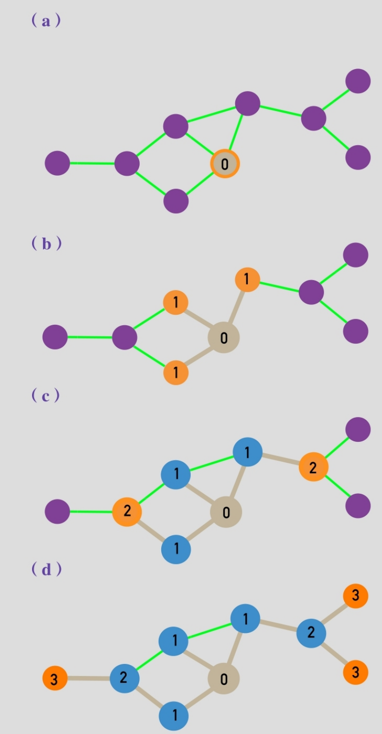


图**1-14**　广度优先搜索算法示例

（**a**）从标记为0的橙色节点出发，找出其所有邻居节点，并标记为1。

（**b**）～（**d**）接下来，对于标记为1的节点，找出其所有未标记的邻居节点，将其标记为2。在随后的迭代中，逐渐增加标记数，直到所有节点都被标记完为止。算法结束后，节点0和任意其他节点i 之间的最短路径长度或者距离d_{0i} 为节点i 的标记。例如，节点0和距离其最远的节点之间的距离为$d=3$。

（**1**）从节点i 出发，将其标记为0；

（**2**） 找出与节点i 直接相连的节点，标记为距离1，然后将它们放到一个队列中；

（**3**） 从队列中取出最前面的节点（其距离标记假设为n ），找出和该节点相连且尚未被标记的节点，将它们标记为$n+1$，然后把它们放在队列的后面；

（**4**） 重复步骤3，直到碰到目标节点j 或者队列中没有节点了；

（**5**）节点i 和节点j 之间的距离为节点j 的标记。如果节点j 没有标记，则$d_{ij}=\infty$

对于有N 个节点和L 个链接的网络而言，BFS算法的计算复杂度（即计算机计算d_{ij} 所需要的步数）为$O(N+L)$。计算复杂度和N 、L 是线性关系，因为每个节点需要进入和离开队列各一次，而每个链接需要被检测一次。

1.9 连通性

手机作为一种通信设备，如果不能呼叫任何有效的手机号码，其用途就非常有限了；电子邮箱如果只能给少数几个邮件地址发邮件，也就没有多大用处了。从网络的角度看，这意味着手机或互联网背后的网络必须能够在任意两个节点之间建立一条路径。实际上，这是大多数网络的关键效用——确保连通性。本节我们将从图论的角度探讨连通性。

无向网络中，如果节点i 和节点j 之间存在一条路径，则它们是连通的。如果不存在路径（$d_{ij}=\infty$），则它们是不连通的。图**1-15a** 所示的网络包含两个互不连通的簇。同一个簇中的任意两个节点（例如节点4和节点6）之间存在路径，不同簇的节点（例如节点1和节点6）之间没有路径。

如果一个网络的所有节点对都是连通的，我们说该网络是连通的。如果网络中至少有一对节点是不连通的，则该网络是不连通的。很明显，图**1-15a** 所示的网络是不连通的，我们把互不连通的两个子网络称为连通分支或簇。一个连通分支是网络节点的子集，该子集满足“任意两个节点之间存在路径”的性质，且在不破坏该性质的前提下无法再增加任何节点。

如果网络中包含两个连通分支，只需要选择合适的位置放置一个链接，就能使网络变成连通的（图**1-15b** ）。这样的链接被称为桥。一般而言，桥是指断开之后网络就不再保持连通的链接。

对于小网络而言，目测就可以帮助我们判断其是否是连通的。对于有数百万个节点的网络而言，判断其连通性则是一个有挑战性的问题。不过，数学工具和计算机算法可以帮助我们找出网络中的连通分支。例如，对于不连通的网络，其邻接矩阵可以通过重排的方式变成分块对角阵的形式——所有非零元素形成沿对角线排列的方块（图**1-15a** ）。每一个方块对应着一个连通分支。我们可以使用线性代数工具来判定邻接矩阵是否是分块对角阵，从而找出网络的连通分支。

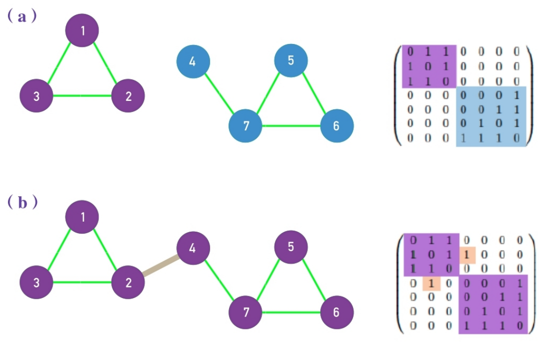


图**1-15** 连通网络和不连通网络

（**a**） 包含两个连通分支的小网络。实际上，第一个连通分支（1,2,3）中的任意节点对之间都存在路径，第二个连通分支（4,5,6,7）也一样。然而，不同连通分支的节点之间不存在路径。

图的右边部分是该网络的邻接矩阵。如果网络是不连通的，则邻接矩阵可以通过重排变成一个分块对角阵——所有非零元素形成沿对角线排列的方块。

（**b**） 增加一个链接（灰色）使网络从不连通变成连通的，这样的链接通常被称为桥。添加该链接之后，网络中每一对节点之间都存在路径了。因此，其邻接矩阵不能写成分块对角阵的形式了。

实际上，对于大网络而言，找出其连通分支的更高效的方法是BFS算法（边栏**1.6**）。

边栏1.6

寻找网络的连通分支

1.从随机选择的一个节点i出发，执行边栏**1.5**描述的BFS算法。将从节点i出发可到达的所有节点标记为$n=1$。

2.如果已标记节点的个数等于N，则网络是连通的。如果已标记节点的个数小于N，则网络包含多个连通分支。找出这些连通分支需要继续执行步骤3。

3.标号增加1，即$n\to n+1$。选择一个未标记的节点j，将其标记为n。使用BFS算法找出从节点j出发可到达的所有节点，将它们都标记为n。重复步骤3，直到标记完所有的节点为止。

1.10 集聚系数

集聚系数刻画了一个节点的邻居节点之间彼此连接的稠密程度。对于一个度为k_i的节点i，其局部集聚系数定义为[12]：

$$C_i=\frac{2L_i}{k_i(k_i-1)} \tag{1.15}$$

这里的L_i表示节点i的k_i个邻居之间的链接数。

注意，集聚系数的取值介于0和1之间（图**1-16a**）：

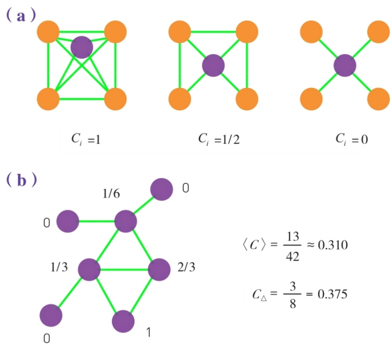


图1-16　集聚系数

（**a**）度为k_i 的中心节点i 在三种不同连接情况下的局部集聚系数C_i 。局部集聚系数刻画了节点的局部链接密度。

（**b**）　网络集聚系数的示例网络。每个节点的局部集聚系数展示在了节点的附近。另外，图中还给出了网络平均集聚系数$\langle C \rangle$ 和全局集聚系数$C_{\triangle}$ ——**1.13**节 公式**1.17** 。注意，对于度为0或1的节点，集聚系数为0。

- C_i ＝0表示节点i 的所有邻居彼此都不相连。
- C_i ＝1表示节点i 的所有邻居形成了一个完全图，即邻居两两相连。
- C_i 表示节点i 的任意两个邻居彼此相连的概率。因此，C_i ＝0.5意味着节点i 的任意两个邻居有50%的概率彼此相连。

总之，C_i 刻画了网络的局部链接密度：节点i 的邻居之间连接越紧密，其局部集聚系数就越高。

整个网络的集聚程度可以通过其所有节点的平均集聚系数$\langle C \rangle$ 来刻画，定义为

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i \qquad (1.16)$$

平均集聚系数$\langle C \rangle$ 的概率化解释是：随机选择一个节点，其两个邻居彼此相连的概率。

公式**1.16** 是在无向网络上集聚系数的定义，集聚系数的定义同

样可以扩展到有向网络和加权网络上[13]，[14]，[15]，[16]。在网络科学文献中，我们可能还会看到全局集聚系数的概念（进阶阅读**1.A**）。

1.11 小结

本节简要介绍了图论的一些基本概念和网络科学使用的一些工具。图**1-17**总结了网络的一些基本性质，为我们进一步研究网络提供了形式化语言。

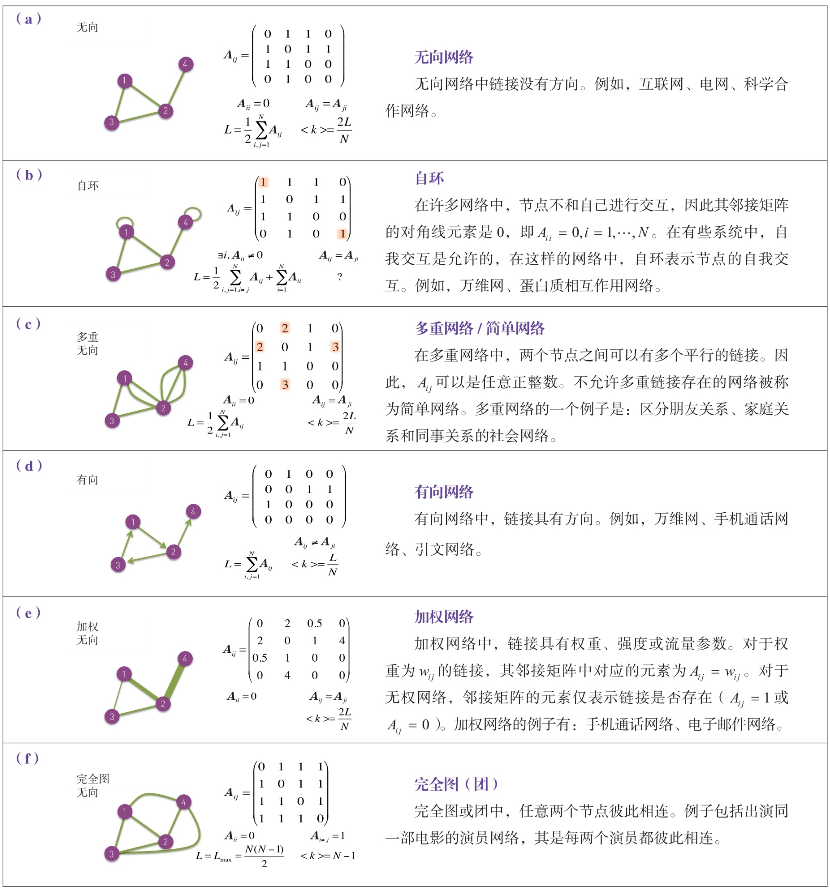


图1-17　图表学

在网络科学领域，我们经常按照网络的一些基本性质对其进行区分。这里，我们将概述最常见的一些网络类型，还列出了具有某个特定性质的真实系统。注意，很多真实网络可以同时具备这些基本性质中的多个。例如，万维网是带自环的无向多重图；手机通话网络是有向加权无自环的网络。

网络科学研究的许多网络包含数千个甚至数百万个节点和链接（表**1-1**）。因此，除了图**1-17** 所示的那些小网络，我们还需要走得更远。如果想初步了解我们将会碰到的网络，可以看一下图**1-4a** 所示的酵母蛋白质相互作用网络。该网络过于复杂，目测的方式对于理解和认识这类网络不再适用。因此，我们需要使用网络科学的工具来刻画网络的拓扑。

首先，我们使用本节介绍的基本方法来测量蛋白质相互作用网络的特性。如图**1-4a** 所示，该网络是一个无向网络，有$N=2\ 018$个蛋白质作为节点，$L=2\ 930$个交互关系作为链接。因此，根据公式**1.2**，该网络的平均度为$\langle k\rangle=2.90$，表明一个蛋白质和两到三个其他蛋白质存在相互作用。不过，这个数字是有误导性的。实际上，如图**1-4b**、**c** 中的度分布p_k所示，大多数节点只有很少几个链接。精确地讲，该网络中69%的节点有少于3个链接，即$k<\langle k\rangle$。大量只有少数几个链接的节点和少数几个有很多链接的枢纽节点并存——最大的枢纽节点有多达92个链接。节点的度差异之大，源于第**3**章 介绍的网络无标度特性。我们将看到，度分布的形状决定着网络的很多性质，例如网络健壮性和病毒的传播。

其次，广度优先搜索算法（边栏**1.5**）可以帮助我们确定该网络直径为$d_{\max}=14$。我们可能会认为节点间距离d的差异也会很大，因为有些节点距离很近，其他节点则相距很远。然而，如图**1-18a** 所示，距离的分布却不是这样的：距离的分布p_d在5和6之间有一个峰值，意味着大部分节点间的距离比较小，在平均距离$\langle d\rangle=5.61$附近。另外，随着d的增加，p_d快速减小，这意味着大的距离很少出现。实际上，距离分布的方差是$\sigma_d=1.64$，这说明大部分路径的长度在平均值$\langle d\rangle$的附近。这就是第**2**章 讨论的小世界性质的体现。

广度优先搜索算法还告诉我们，蛋白质相互作用网络不是连通的，而是由185个连通分支组成，参见图**1-4a** 中孤立的簇和节点。被

称为巨连通分支的最大连通分支包含了全部2 018个节点中的1 647个；其他所有连通分支都很小。我们在随后的章节中会看到，这是真实网络中的常见情形。

蛋白质相互作用网络的平均集聚系数是$\langle C\rangle = 0.12$。我们在随后的章节中会看到，这意味着高度的局部聚集性。另一个启示来自集聚系数和节点度之间的关系，即图**1-18b** 所示的函数$C(k)$。$C(k)$随着节点度k的增加而减小，这表明小度节点的局部集聚系数明显高于枢纽节点的局部集聚系数。因此，小度节点的邻居之间连接相对稠密，而枢纽节点的邻居之间连接相对稀疏。这是第**8**章 讨论的网络层级结构的必然结果。

最后，还有一个有趣的现象：枢纽节点倾向于和小度节点相连，使得该网络呈现出星型的轮辐结构（图**1-4a**）。这源于第**6**章 探讨的度相关性。度相关性影响着很多基于网络的过程，包括信息扩散过程、控制网络所需的驱动节点个数等。

综上所述，如图**1-4** 和图**1-18** 所展示的那样，本章介绍的一些量可以帮助我们认识真实网络的一些关键性质。随后的章节将系统地研究这些网络性质，带你理解这些网络性质在认识复杂系统方面发挥的重要作用。

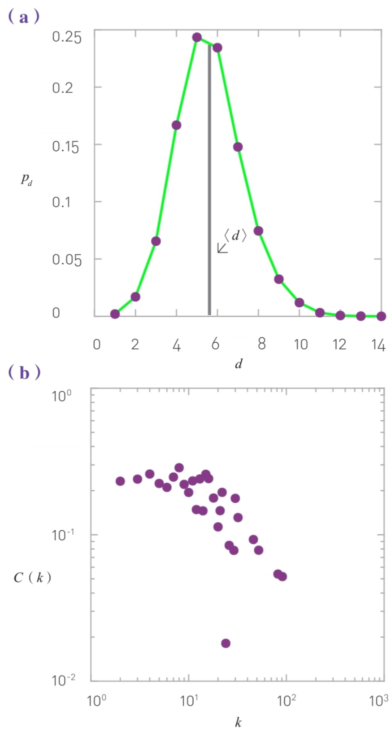


图1-18　刻画真实网络

酵母蛋白质相互作用网络经常被生物学家和网络科学家研究。该网络的拓扑如图**1-4a** 所示，网络中包含$N=2\ 018$个节点和$L=2\ 930$个链接，其中一个连通分支占据了全部节点的81%，除此之外，还有一些小的连通分支和许多孤立的节点。

（**a**）　蛋白质相互作用网络的距离分布p_d 反映了随机选择的两个节点之间距离为d（最短路径的长度）的概率。图中灰色的垂直线对应着平均路径长度$\langle d\rangle=5.61$。

（**b**）平均局部集聚系数和节点度的关系。函数$C(k)$是度为k 的节点的平均局部集聚系数。

1.12 课后习题

习题1　哥尼斯堡七桥问题

图1-19 中哪个图可以一笔画成（要求每条线只能画一次）？并给出理由。

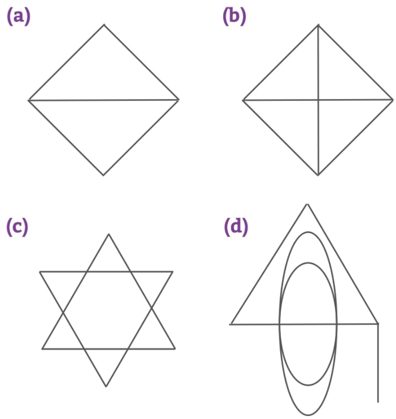


图1-19 哥尼斯堡七桥问题

习题2 矩阵运算

令A 表示一个无向无权无自环网络的邻接矩阵，1是一个有N 个元素且每个元素都为1的列向量。换句话说，$1=(1,1,\ldots,1)^T$ ，这里的T 表示矩阵转置。使用矩阵运算（数乘、行向量乘列向量、矩阵转置、矩阵的迹等，不能使用求和符号）写出如下表达式：

（1）元素是各个节点的度k_i $(i=1,2,\ldots,N)$的向量k。

（2）网络中链接总数L 。

（3）网络中三角形个数T 。三角形是指两两相连的三个节点（提示：可以使用矩阵的迹）。

（4）向量k_{nn} ：第i 个元素为节点i 的邻居节点的度之和。

（5）向量k_{nnn} ：第i 个元素为节点i 的二阶邻居（邻居的邻居）的度之和。

习题3 网络表示

在许多分析运算中，邻接矩阵是一种非常有用的网络表示方式。不过，在计算机中存储网络时，使用邻接表的方式可以节省存储空间。邻接表是一个大小为$L \times 2$的矩阵，每行对应一个链接，记录该链接的起始节点和终止节点。对于图**1-20** 中的两个网络：

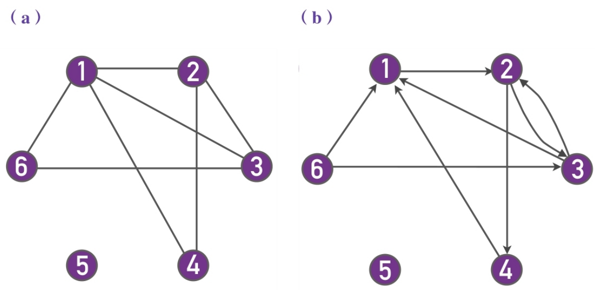


图**1-20**　网络表示

（**a**）有6个节点和7个链接的无向网络；

（**b**）有6个节点和8个有向链接的有向网络。

（**1**）写出其邻接矩阵；

（**2**）写出其邻接表；

（**3**）计算图**1-20a** 所示网络的平均集聚系数；

（**4**）交换图**1-20a** 中标记为5和6的节点，看看其邻接矩阵和邻接表的变化；

（**5**）关于网络的信息，有哪些是可以从其邻接矩阵表示中得到而无法从其邻接表表示中得到的？

（**6**）在网络（a）中，从节点1出发到节点3，有多少条长度为3的路径（节点和链接可以重复）？网络（b）呢？

（**7**）编写程序计算两个网络中长度为3的环的个数。

习题4　度、集聚系数和连通分支

（**1**）假如一个无向网络含有N个节点，所有节点的度都为1。那

么，节点个数N 需要满足什么条件？该网络的度分布如何？网络包含多少个连通分支？

（2） 假如一个无向网络含有N 个节点，所有节点的度都为2，集聚系数都为1。那么，该网络的拓扑结构是什么样的？节点个数N 需要满足什么条件？

习题5　二分网络

对于图1-21 所示的二分网络，

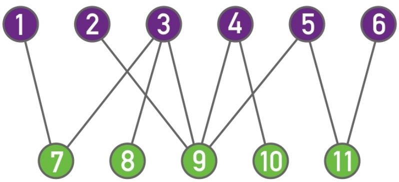


图1-21　二分网络

该二分网络中，一个节点集合包含6个节点，另一个节点集合包含5个节点，两个节点集合之间通过10个链接相连。

（1） 写出其邻接矩阵，并说明其邻接矩阵为什么是分块矩阵。

（2） 将该二分网络分别映射到紫色节点和绿色节点上，并分别写出两个映射网络的邻接矩阵；

（3） 分别计算紫色节点和绿色节点的平均度；

（4） 对于（3）中得到的两个映射网络，分别计算其节点的平均度。

习题6　一般意义的二分网络

假设二分网络的两个节点集合的大小分别是N_1 和N_2。

（1） 写出该二分网络的最大可能链接数$L_{\max}$；

（2） 和大小为N_1+N_2 的非二分网络相比，有多少个链接不能

出现在该二分网络中？

（3）假如$N_1 \ll N_2$，对于该二分网络的链接密度——实际链接数和最大链接数L_{max}的比值，有哪些发现？

（4）分别写出两个节点集合的平均度$\langle k_1 \rangle$和$\langle k_2 \rangle$，并说明二者之间的联系。

1.13 进阶阅读1.A 全局集聚系数

在网络科学的文献中，我们偶尔会碰到全局集聚系数的概念。全局集聚系数用于计算网络中闭合三角形的总数。实际上，公式**1.15**中的L_i是节点i参与的三角形个数：节点i的邻居节点之间的每个链接对应着节点i参与的一个闭合三角形（图**1-17**）。因此，网络的全局集聚程度可以通过全局集聚系数来体现，全局集聚系数定义为：

$$C_{\Delta} = \frac{3 \times \text{三角形个数}}{\text{连通三元组的个数}} \quad (1.17)$$

这里的连通三元组是由3个节点构成的集合（假设为ABC），要求节点A和节点B之间存在链接，节点B和节点C之间存在链接。例如，节点A、B和C构成的一个三角形包含3个这样的连通三元组，分别是ABC、BCA和CAB。假如A和B之间有链接、B和C之间有链接、A和C之间没有链接，那么节点A、B和C形成一个非闭合三角形。公式**1.17**分子中的因子为3是由于每个三角形贡献3个连通三元组。全局集聚系数的概念起源于20世纪40年代的社会网络文献[17]，[18]，在这些文献中，C_{Δ}通常被称为传递三元组比例。

注意，公式**1.16**中的平均集聚系数和公式**1.17**中的全局集聚系数不等价。以包含N个节点的双中心星型网络为例，网络的两个中心节点——节点1和节点2彼此相连，且均与网络中的其他节点彼此相连，而其他节点之间没有链接。该网络中，节点1和节点2的局部集聚

系数为2 /（N -1），而其他节点的集聚系数均为1，因此平均集聚系数为$\langle C\rangle$ =1-O （1）。然而，该网络的全局集聚系数为$C_\Delta\sim1/N$ ，和平均集聚系数相差很大。多数情况下，二者之间的差异不像这个极端情形下那么大，但仍然不相同 [19] 。例如，对于图**1-16b** 所示的网络，平均集聚系数为$\langle C\rangle$ =0.31，而全局集聚系数为C_Δ =0.375。

1.14 参考文献

[1] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *PNAS* , 104:8685-8690, 2007.

[2] H.U. Obrist. *Mapping it out: An alternative atlas of contemporary cartographies* . Thames and Hudson, London, 2014.

[3] I. Meirelles. *Design for Information* . Rockport, 2013.

[4] K. Börner. *Atlas of Science: Visualizing What We Know* .The MIT Press, 2010.

[5] L. B. Larsen. *Networks: Documents of Contemporary Art* .MIT Press. 2014.

[6] L. Euler, Solutio Problemat is ad Geometriam Situs Pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8:128-140, 1741.

[7] G. Alexanderson. Euler and Königsberg's bridges: a historical view. *Bulletin of the American Mathematical Society* 43: 567, 2006.

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[9] G. Gilder. Metcalfe's law and legacy. *Forbes ASAP* , 1993.

[10] B. Briscoe, A. Odlyzko, and B. Tilly. Metcalfe's law is wrong. *IEEE Spectrum* , 43:34-39, 2006.

[11] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási. Flavor network and the principles of food pairing, *Scientific Reports* , 196, 2011.

[12] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world' networks. *Nature* , 393:440-442, 1998.

[13] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS* , 101:3747-3752, 2004.

[14] J. P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E* , 71:065103, 2005.

[15] B. Zhang and S. Horvath. A general framework for weighted gene coexpression network analysis. *Statistical Applications in Genetics and Molecular Biology* , 4:17, 2005.

[16] P. Holme, S. M. Park, J. B. Kim, and C. R. Edling. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A* , 373:821-830, 2007.

[17] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika* , 14:95-116, 1949.

[18] S. Wasserman and K Faust. *Social Network Analysis:Methods and Applications* . Cambridge University Press, 1994.

[19] B. Bollobás and O. M. Riordan. Mathematical results on scale-free random graphs, in Stefan Bornholdt, Hans Georg Schuster, *Handbook of Graphs and Networks: From the Genome to the Internet* (2003 Wiley-VCH Verlag GmbH & Co. KGaA).

第2章
随机网络

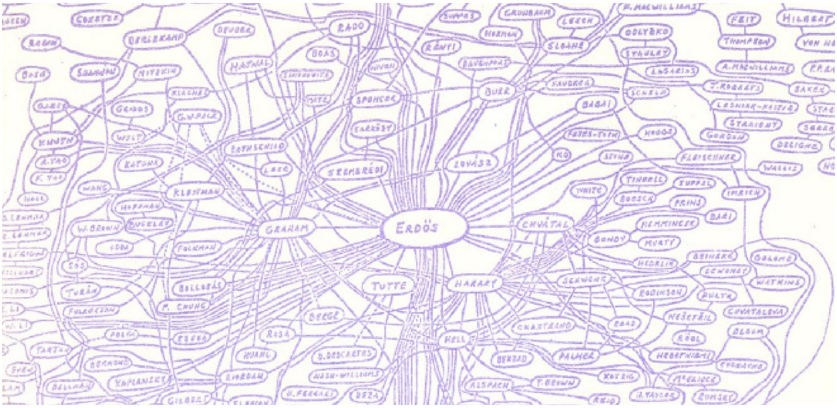


图2-0　埃尔德什数

匈牙利数学家保罗·埃尔德什发表过数百篇研究论文，其中很多论文是和其他数学家合作完成的。他这种坚持不懈的合作研究方式促成了埃尔德什数的产生。埃尔德什数理论是这样的：埃尔德什本人的埃尔德什数是0，与他合作的人的埃尔德什数是1。那些与埃尔德什数为1的人合作写过论文的人，其埃尔德什数是2，以此类推。对于通过合作关系链无法连接到埃尔德什的人，其埃尔德什数记为无穷大。许多著名科学家的埃尔德什数都很低：阿尔伯特·爱因斯坦的埃尔德什数为2，理查德·费曼的埃尔德什数为3。该图由埃尔德什的亲密合作者之一罗纳德·格雷厄姆（Ronald Graham）在1970年绘制，图中展示了埃尔德什的合作者。随着埃尔德什名气的攀升，这幅图也取得了标志性地位。

本章索引

2.1 本章概览 / 073

2.2 随机网络模型 / 075

2.3 链接数 / 078

2.4 度分布 / 081

2.5 真实网络不是泊松分布的 / 084

2.6 随机网络的演化 / 087

2.7 真实网络是超临界的 / 093

2.8 小世界 / 095

2.9 集聚系数 / 101

2.10 小结：真实网络不是随机的 / 104

2.11 课后习题 / 108

2.12 进阶阅读2.A
推导泊松分布 / 111

2.13 进阶阅读2.B
最大度和最小度 / 113

2.14 进阶阅读2.C
巨连通分支 / 115

2.15 进阶阅读2.D
分支大小 / 117

2.16 进阶阅读2.E
全连通状态 / 119

2.17 进阶阅读2.F
相变 / 120

2.18 进阶阅读2.G

小世界的修正 / **122**

2.19 参考文献 / **124**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森 **SARAH MORRISON**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

阿迈勒·侯赛尼 **AMAL HUSSEINI**

莫罗·马蒂诺 **MAURO MARTINO**

菲利普·赫费尔 **PHILIPP HOEVEL**

罗伯塔·西纳特拉 **ROBERTA SINATRA**

2.1 本章概览

让我们想象这样一个情景：为100位最初互不相识的客人组织的一场酒会。[1] 酒会为客人们提供了酒水和奶酪，你很快就会看到客人们三两成群地开始交谈。此时，你提醒其中一位客人玛丽，告诉她那瓶没有标签的深绿色瓶子里的红葡萄酒是罕见的酒中佳品，比那瓶带着精致红色标签的酒要好得多。如果她只与熟人分享该信息，那么你那瓶昂贵的红葡萄酒看似是安全的，因为酒会期间她只会接触到少数几个熟人。

不过，客人们会不停地走动交谈，在原本彼此陌生的人之间形成微妙的“熟人”路径。例如，在约翰遇到玛丽之前，他们二人都遇到了迈克，因而通过迈克形成了一条从约翰到玛丽的无形路径。随着时间的推移，客人们将会通过这样的无形路径彼此交织在一起。如此一来，关于那瓶无标签葡萄酒的秘密便会从玛丽传到迈克，又从迈克传到约翰，进而迅速扩散到更大的群体中（图**2-1** ）。

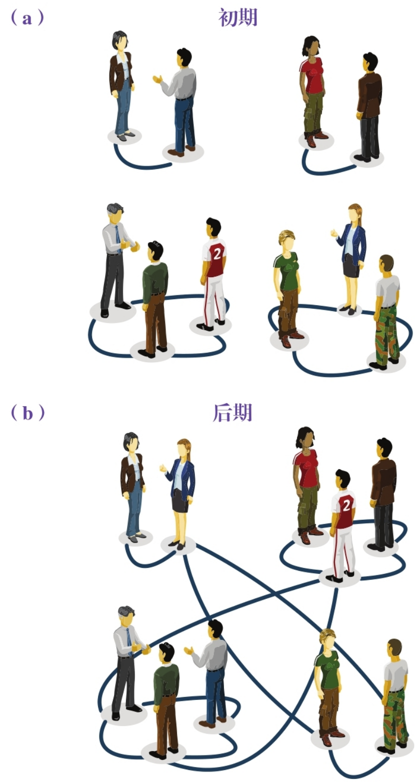


图2-1　从鸡尾酒会到随机网络

鸡尾酒会上通过随机相遇形成的“熟人”网络：

（a）酒会初期，客人们形成孤立的群组。

（b）　酒会后期，随着客人们的走动，群组不断发生变化，一个将所有客人连接在一起的无形网络形成了。

可以肯定的是，当所有客人都彼此认识之后，每个人都会去倒那瓶好葡萄酒。假设每次相遇花费10分钟，一个人遇到所有其他99个人将需要花费大约6小时的时间。因此，你有理由相信，客人们走之后，那瓶好葡萄酒还会剩下一部分供你自己享用。

不过，你要是那样想就错了。本章内容将会向你展示为什么会这样。在这一章中，我们将会看到，这个聚会问题对应着网络科学中的一个经典模型——随机网络模型。随机网络理论告诉我们，不需要等

到所有人都彼此认识，你那瓶昂贵的葡萄酒就已经危险了。实际上，在每个客人遇到至少一个其他客人之后不久，客人们之间就会形成一个无形的网络，借助该网络，关于那瓶葡萄酒的信息很快就能够传到每个人那里。因此，很快每个人都会开始品尝那瓶更好的葡萄酒了。

2.2 随机网络模型

网络科学旨在建立能够重现真实网络性质的模型。我们遇到的大多数真实网络没有晶格所具有的那种令人愉悦的规则结构，或者蜘蛛网所具有的那种可预测的放射型结构。相反，真实网络乍一看好像是随机连接而成的（图**1-4**）。随机网络理论通过构建和刻画真正随机的网络来解释这种表面上的随机性。

从建模的角度来看，网络是一个相对简单的对象，仅由节点和链接组成。然而，真正的挑战在于，在哪些节点间放置链接才能重现真实系统的复杂性。在这一点上，随机网络背后的哲学思想很简单：在节点之间随机放置链接。因此，随机网络定义如下（边栏**2.1**）：

边栏2.1

定义随机网络

随机网络有两种定义方式：

***G*（*N*，*L*）模型**

N 个节点通过L 条随机放置的链接彼此相连。埃尔德什和雷尼在他们关于随机网络的系列论文[2-9]中采用的是这种定义方式。

***G*（*N*，*p*）模型**

N 个节点中，每对节点之间以概率p 彼此相连。该模型是由埃德加·N.吉尔伯特（Edgar N. Gilbert）提出的。[10]

G（N，p）模型固定了两个节点的连接概率p，G（N，L）

模型则固定了总链接数L。$G(N, L)$模型中，节点平均度可以简单地算出，即$\langle k \rangle = 2L/N$。$G(N, p)$模型中，网络其他特征则更容易计算。本书将主要探讨$G(N, p)$模型，不仅是因为该模型便于计算网络的一些关键特征，还因为在真实网络中，链接数很少保持固定不变。

随机网络由N个节点组成，每对节点相互连接的概率为p。

构建随机网络的步骤如下：

（**1**）从N个孤立节点开始。

（**2**）选择一对节点，产生一个0到1之间的随机数。如果该随机数小于p，在这对节点之间放置一条链接；否则，该节点对保持不连接。

（**3**）对所有$N(N-1)/2$个节点对，重复步骤（2）。

上述过程得到的网络被称为随机图或随机网络。两位数学家——保罗·埃尔德什和阿尔弗雷德·雷尼在理解随机网络的性质方面发挥了重要作用。为了纪念他们，随机网络被命名为埃尔德什 - 雷尼网络（边栏**2.2**）。

边栏**2.2**

随机网络简史

阿纳托尔·拉波波特是一个移民到美国的俄罗斯人，也是第一个研究随机网络的人。拉波波特最初想成为一名钢琴演奏家，在他意识到一名成功的钢琴演奏家需要有丰厚的财力支持后，拉波波特的兴趣转到了数学上。在他专注于数学和生物学的时候，数学家和生物学家之间几乎没什么交流。在其1995年和雷·索洛莫诺夫合作的论文中[11]，拉波波特指出，逐渐增加网络的平均度，网络连接性会出现一个突然的变化，从彼此不连通的节点组变成一幅具有一个巨大连通分支的图。

保罗·埃尔德什和阿尔弗雷德·雷尼（图**2-2**　）在随机网络方面做出的基础性工作使随机网络的研究达到鼎盛。1959年到1968年间发表的8篇论文中 [2-9]　，他们在图论中融入了概率论和组合数学，建立了数学领域的一个新分支——随机图论 [2] 。

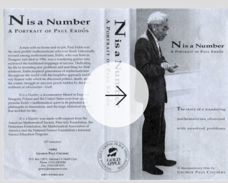


图**2-2**　（**a**）保罗·埃尔德什（**1913—1996**）

匈牙利数学家，因其卓越的科学产出和古怪的个性而闻名。实际上，埃尔德什发表的论文数比数学史上任何其他数学家都要多。他和500多个数学家合作撰写过论文，促成了埃尔德什数概念的产生。他传奇般的个性和深远的影响带来了两部传记 [12] ，[13] 和一部纪录片 [14] （在线资源**2.1**　）。

（**b**）阿尔弗雷德·雷尼（**1921—1970**）

匈牙利数学家，对组合数学、图论和数论都有重要贡献。他的影响超出了数学领域：雷尼熵广泛应用于混沌理论，他参与建立的随机网络理论是网络科学的核心。他的名字被用于命名布达佩斯的阿尔弗雷德·雷尼数学研究院——匈牙利数学的温床。



在线资源**2.1**　《**N**是一个数：保罗·埃尔德什的肖像》

1993年发行的保罗·埃尔德什的传记纪录片，由乔治·保罗·西泽瑞（George　Paul Csicsery）导演。这部纪录片讲述了埃尔德什的生活及其在科学领域的影响 [14] 。

扫描二维码下载“湛庐阅读” App，回复书名，观看保罗·埃尔德什的传记纪

录片。

在埃尔德什和雷尼发表他们第一篇随机网络论文的同一年，埃德加·N.吉尔伯特 [10] 也独立提出了他的随机网络模型。不过，由于埃尔德什和雷尼的研究工作影响巨大，他们二人被认为是随机图论的奠基人。

“数学家是一种把咖啡变成定理的装置。”

——阿尔弗雷德·雷尼

（人们通常认为是埃尔德什所说）

2.3 链接数

由同样的参数N 和p 产生的随机网络，看起来会稍有不同（图**2-3**）。这种不同不仅体现在详细的节点连接情况上，还体现在链接数L 上。因此，给定参数N 和p 时，判定出所生成随机网络的期望链接数是有价值的。

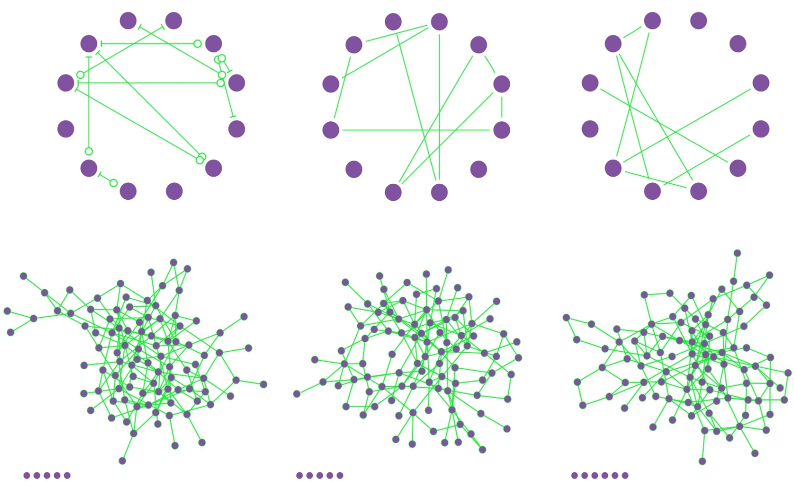


图**2-3**　随机网络是真正随机的

第一行图

参数均为p =1/6，N =12的三个随机网络。尽管参数相同，但这三个网络不仅看上去差别很大，链接数也不同（L =10，10，8）。

第二行图

参数均为p =0.03，N =100的三个随机网络。可以看到，图的底部有一些孤立节点，这些节点的度为k =0。

随机网络恰好有L 条链接的概率，是如下三项的乘积：

（1）L 个点对之间存在链接的概率，即p^L 。

（2） 剩余$N(N-1)/2-L$ 个点对之间没有链接的概率，即$(1-p)^{N(N-1)/2-L}$ 。

（3） 在所有$N(N-1)/2$个点对中选择L 个点对放置链接，所有可能的选择方式数为：

$$\binom{\frac{N(N-1)}{2}}{L} \tag{2.0}$$

因此，随机网络恰好有L 条链接的概率为：

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L} \tag{2.1}$$

公式**2.1** 是一个二项式分布（边栏**2.3** ），因此随机网络的期望链接数为：

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2} \tag{2.2}$$

边栏2.3

二项分布：均值和方差

假设我们掷一枚硬币，则得到正面和反面的概率均为p

＝1/2。假如我们掷N 次硬币，有x 次正面的概率记为p_x ，则p_x 服从二项分布。一般而言，二项分布用于描述有两个可能结果的N 次独立实验中某个结果出现的次数。我们记一种结果出现的概率为p ，另一种结果出现的概率为1-p 。

二项分布的形式为：

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}$$

分布的均值（一阶矩）为：

$$x = \sum_{x=0}^{N} x p_x = Np \qquad (2.4)$$

其二阶矩为：

$$x^2 = \sum_{x=0}^{N} x^2 p_x = p(1-p)N + p^2N^2 \qquad (2.5)$$

因此，二项分布的标准差为：

$$\sigma_x = \left(x^2 - x^2\right)^{\frac{1}{2}} = \left[p(1-p)N\right]^{\frac{1}{2}} \qquad (2.6)$$

在刻画随机网络的性质时，公式**2.4** ～公式**2.6** 经常会被使用。

可以看出，$\langle L\rangle$ 是节点对之间存在链接的概率p 和链接对总数$L_{\max}$ ＝N (N -1)/2（见第**1**章 ）的乘积。

根据公式**2.2** ，我们可以得到随机网络的平均度

$$\langle k\rangle = \frac{2\langle L\rangle}{N} = p(N-1) \qquad (2.3)$$

也就是说，对于节点数为N 的随机网络而言，其平均度$\langle k\rangle$ 是节点对间存在链接的概率p 和一个节点最多可能拥有的链接数(N -1)的乘积。

综上所述，由相同参数N 和p 产生的不同随机网络，其链接数可以不同。链接数的期望值取决于N 和p 。增大p 的值会使随机网络变得更稠密：平均链接数从$\langle L \rangle = 0$线性增加到L_{max} ，节点平均度由$\langle k \rangle = 0$增加到$\langle k \rangle = N-1$。

2.4 度分布

随机网络中，有些节点有许多链接，有些节点只有少量链接，甚至没有链接（图2-3 ）。这种差异可以通过度分布p_k 来刻画，p_k 表示一个随机选择的节点其度为k 的概率。这一节中，我们将推导出随机网络的度分布p_k 并讨论其性质。

二项分布

随机网络中，一个节点i 恰好有k 个链接的概率是下面三项的乘积[15]：

（1）k 个链接出现的概率，即p^k 。

（2）剩下$(N-1-k)$个链接不出现的概率，即$(1-p)^{N-1-k}$ 。

（3）节点i 的$N-1$个可能存在的链接中选出k 个，选择方式的总数为：

$$\binom{N-1}{k}$$

因此，随机网络的度分布服从二项分布：

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \quad (2.7)$$

二项分布的形状取决于网络大小N 和链接概率p （图2-4 ）。根据二项分布（边栏2.3 ）的性质，我们可以计算出网络的平均度$\langle k \rangle$——公式2.3 ，也可以计算出度分布的二阶矩$\langle k^2 \rangle$ 和方差σ_k （图2-4

）。

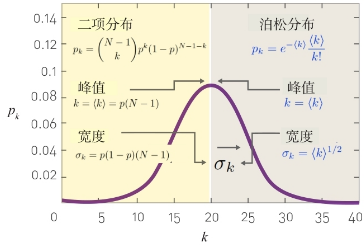


图2-4　二项度分布和泊松度分布

随机网络度分布的精确形式是二项分布（左半部分）。当$N \gg \langle k \rangle$时，二项分布可以很好地近似为泊松分布（右半部分）。由于二者在描述同一个分布，它们有着相同的性质，只不过表述这些性质的参数不同：二项分布取决于p和N，而泊松分布只有一个参数$\langle k \rangle$。正是泊松分布的这种简单性使其更受青睐。

泊松分布

大部分真实网络是稀疏的，意味着这些网络的平均度远小于网络大小——$\langle k \rangle \ll N$（表**1-1**）。极限情况下，公式**2.7**中的度分布可以近似为如下泊松分布（进阶阅读**2.A**）

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (2.8)$$

公式**2.8**和公式**2.7**通常被称为随机网络的度分布。

公式**2.7**中的二项分布和公式**2.8**中的泊松分布是在描述同一个量，因此它们有相似的性质（图**2-4**）：

- 这两个分布都在$\langle k \rangle$附近有一个峰值。增加p的值会使网络变得稠密，平均度$\langle k \rangle$和度分布的峰值会右移。
- 分布的宽度（离散度）也由p和$\langle k \rangle$控制。网络越稠密，分布越宽，节点度的差异也越大。

使用泊松形式（公式**2.8**）的度分布时需要注意：

- 随机网络度分布的精确形式是二项分布（公式**2.7**）。因

此，公式2.8 只是在$\langle k \rangle \ll N$ 的极限情况下对公式2.7 的近似。由于大多数真实网络都是稀疏的，上述近似所需的条件通常会被满足。

- 泊松形式的优势是，像$\langle k \rangle$、$\langle k^2 \rangle$和σ_k 等网络关键特性的形式更简单（图2-4 ），仅依赖于$\langle k \rangle$这一个参数。
- 公式2.8 中的泊松分布不显式地依赖节点数目N 。因此，公式2.8 告诉我们，平均度$\langle k \rangle$相同但大小不同的随机网络，其度分布几乎一样。

综上所述，虽然泊松分布只是随机网络度分布的一种近似，但其形式简单，便于分析，因此在刻画随机网络的度分布p_k 时，人们更倾向于使用泊松形式。本书中，除非特别指出，我们将使用泊松形式（公式2.8 ）来刻画随机网络的度分布。泊松形式度分布的一个关键特征是，其性质与网络大小无关，仅依赖于平均度$\langle k \rangle$ 这一个参数（图2-5 ）。

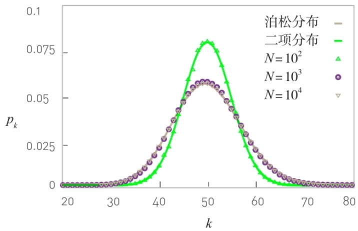


图2-5　度分布与网络大小无关

平均度$\langle k \rangle=50$，大小分别为$N=10^2$、$N=10^3$、$N=10^4$ 的三个随机网络的度分布。

小网络：二项分布

对于小网络（$N=10^2$ ），由于不满足泊松近似的条件$N \gg \langle k \rangle$，该网络的度分布明显偏离泊松分布（公式2.8 ）。因此，小网络的度分布需要使用精确的二项分布形式（公式2.7 ）（绿线）。

大网络：泊松分布

对于大网络（$N=10^3$，$N=10^4$ ），其度分布与灰线所示的泊松分布（公式2.8 ）相差无几。因此，当网络大小N 很大时，度分布和网络大小无关。为了避免随机性带来的噪声，图中所示的结果是在1 000个独立生成的随机网络上平均得到的。

2.5
真实网络不是泊松分布的

随机网络中节点的度可以在0和$N-1$之间变化。由此引发的一个问题是，随机网络中节点度之间的差异有多大？或者说，大度节点与小度节点可以同时存在吗？本节，我们将通过估算随机网络中的最大节点度和最小节点度来回答上述问题。

假设我们的社交网络可以由随机网络模型来描述。这样一个随机社会也不是那么遥不可及：在这个社会中，我们会遇见谁以及和谁成为熟人都充满了随机性。

据社会学家估计，一个普通人大约认识1 000个人。因此，我们假设网络的平均度为$\langle k\rangle \approx 1\ 000$。根据我们前面掌握的关于随机网络的知识，对于拥有$N \approx 7\times 10^9$个人的随机社会，我们可以得出如下一些结论（进阶阅读**2.B**）：

（**1**）随机社会中连通性最好的人（度最大的节点），大约有$k_{max}=1\ 185$个熟人。

（**2**）连通性最差的人，认识大约$k_{min}=816$个人，和k_{max}或$\langle k\rangle$的差异并不大。

（**3**）随机网络度分布的标准差为$\sigma_k=\langle k\rangle^{1/2}$。对于平均度$\langle k\rangle \approx 1\ 000$的随机网络，其标准差为$\sigma_k=31.62$。这意味着一个普通人认识的人数在$\langle k\rangle \pm \sigma_k$的范围内，即介于968和1 032之间非常狭窄的范围。

总之，在随机社会中，每个人的朋友数大体相当。因此，如果我们是随机连接在一起的，就不会有异常的人存在：没有非常受欢迎的人，也没有只有少数几个朋友的人。这一惊人的结论源于随机网络的一个重要性质：在大的随机网络中，大多数节点的度分布在$\langle k\rangle$附近狭窄的范围内（边栏**2.4**）。

边栏**2.4**

为什么没有枢纽节点？

为了理解随机网络中为什么没有枢纽节点——度非常大的节点，我们来看一下度分布（公式**2.8**）。

首先，公式**2.8** 中的$1/k!$项大大降低了大度节点出现的概率。根据斯特林近似（Stirling approximation），有：

$$k! \sim \left[\sqrt{2\pi k}\right]\left(\frac{k}{e}\right)^{k}$$

因此，公式**2.8** 可以重写为：

$$p_k = \frac{e^{-\langle k\rangle}}{\sqrt{2\pi k}}\left(\frac{e\langle k\rangle}{k}\right)^{k} \qquad (2.9)$$

对于度$k > e\langle k\rangle$的节点，括号中的项小于1。因此，对于大的k，公式**2.9** 中与k相关的项，例如$1/\sqrt{k}$和$(e\langle k\rangle/k)^k$，随着k的增加迅速减小。总之，公式**2.9** 表明，随机网络中观测到枢纽节点的概率比指数下降得还要快。

上述结论明显和现实不符。实际上，很多人认识的人数远远超过1185。例如，美国总统富兰克林·德拉诺·罗斯福（Franklin Delano Roosevelt）的预约本中有22 000个名字，这些都是他要亲自会见的人[16]，[17]。与之类似，对Facebook社交网络的一项研究表明，有很多人的好友数达到了5 000——5 000是Facebook社交网络平台设置的好友数最大值[18]。要弄清楚前述结论和现实之间这些巨大差异产生的原因，我们需要对比真实网络的度分布和随机网络的度分布。

图**2-6** 展示了三个真实网络的度分布以及相应的泊松拟合。该图展现了真实网络和随机网络之间的巨大差异：

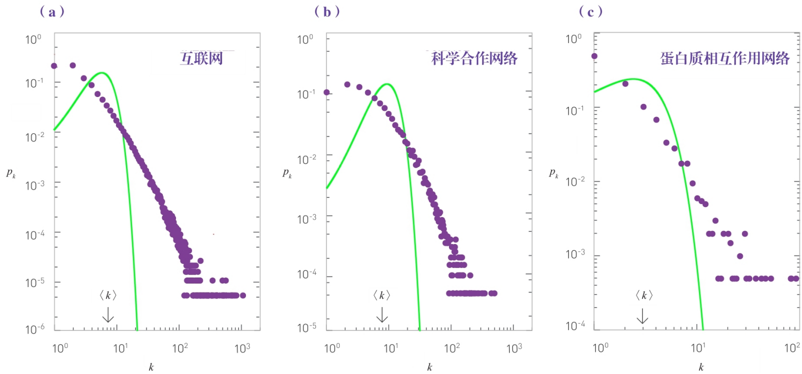


图2-6　真实网络的度分布

度分布：（a）互联网；（b）科学合作网络；（c）蛋白质相互作用网络（表**1-1**）。绿色的线是泊松分布——通过测量真实网络的平均度$\langle k\rangle$　而绘制（参见公式**2.8**）。真实数据和泊松分布之间的显著差异表明，随机网络模型低估了大度节点的度和出现概率，也低估了小度节点的数量。相反，随机网络模型预测网络中大量节点的度在平均度$\langle k\rangle$附近。

（**1**）　泊松形式明显低估了大度节点的个数。例如，根据随机网络模型，互联网的最大度预计在20左右。真实数据表明，有的路由器的度可以接近10^3。

（**2**）　真实网络中度的分布范围比随机网络所预计的要宽得多。这种差异可以通过图**2-4**　所示的分布离散度σ_k　看出。如果互联网是随机的，则预计σ_k　=2.52。而真实测量结果表明，σ_{internet}　=14.14，明显高于随机网络的预计值。这些差异不仅存在于图**2-6**　所示的网络，表**1-1**中列出的所有网络都具有该性质。

总之，和真实数据的对比表明，随机网络模型不能刻画出真实网络的度分布。随机网络中，大多数节点都有类似的度，不存在枢纽节点。与之相反，在真实网络中，我们观察到很多高度连接的节点，节点的度之间有很大的差异。我们将在第**3**章　来解答这些差异产生的原因。

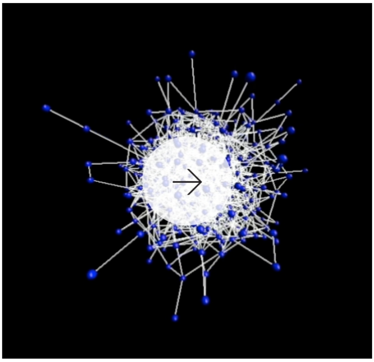
2.6

随机网络的演化

本章开始时介绍的鸡尾酒会刻画了这样一个动态过程：从N 个孤立节点开始，链接逐个放置在随机相遇的客人之间。这一过程等价于逐渐增加p 的值，从而对网络拓扑结构产生影响（在线资源**2.2**）。为了量化这一过程，我们首先看一看网络中最大连通分支的大小N_G是如何随着$\langle k\rangle$而变化的。两个容易理解的极端情形如下：

- 当$p=0$时，$\langle k\rangle=0$，所有节点都是孤立的。因此，最大连通分支的大小为$N_G=1$。对于大的N，有$N_G/N\rightarrow 0$。

- 当$p=1$时，$\langle k\rangle=N-1$，网络是完全连通图，所有节点属于同一个连通分支。因此，$N_G=N$，$N_G/N=1$。



在线资源**2.2**　随机网络的演化

该视频展示了随机网络的结构随着p 的增加而变化的情况。从视频中可以形象地看出，当p 很小时，网络中没有巨连通分支，而当p 到达临界值时，一个巨连通分支突然就出现了。

扫描二维码下载“湛庐阅读”App，回复书名，观看随机网络的演化过程。

有人可能会认为，随着平均度$\langle k \rangle$从0增加到N -1，最大连通分支的大小是逐渐由N_G =1变化到N_G =N 的。然而，如图**2-7a** 所示，情况并非如此：当$\langle k \rangle$比较小时，N_G /N 一直为0，意味着没有大的连通分支出现；一旦$\langle k \rangle$超过临界值，N_G /N 开始迅速增加，标志着一个大的连通分支在快速形成，我们称该连通分支为巨连通分支。埃尔德什和雷尼在他们1959年发表的经典论文中预测了巨连通分支出现的条件[2]：

$$\langle k \rangle = 1 \quad (2.10)$$

换句话说，当且仅当每个节点平均拥有不少于一个链接时，巨连通分支才会出现（进阶阅读**2.C** ）。

每个节点至少需要一个链接才能使网络中出现巨连通分支，这一事实并不让人感到意外。实际上，巨连通分支的存在，要求其所包含的每个节点必须至少和一个该连通分支中的其他节点相连。巨连通分支的出现只需要每个节点拥有一个链接就足够了，这多少有点反直觉，但事实的确如此。

根据公式**2.3** ，可以使用链接概率p 来表示公式**2.10** ：

$$p_c = \frac{1}{N-1} \approx \frac{1}{N} \quad (2.11)$$

因此，网络越大，形成巨连通分支所需要的p 越小。

巨连通分支的出现，只是我们通过改变$\langle k \rangle$ 来刻画随机网络变化时观测到的其中一个转变。实际上，随着$\langle k \rangle$ 变化，我们可以观察到4个拓扑上有明显区别的状态（图**2-7a** ），每个都有其独有的特性：

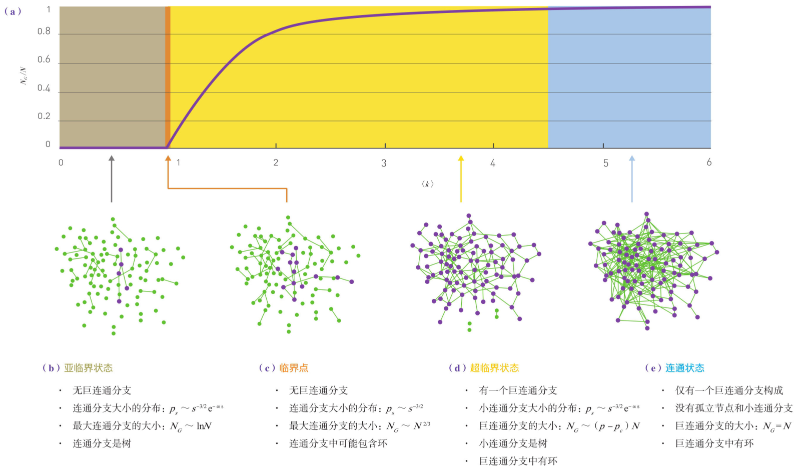


图2-7　随机网络的演化

（a）埃尔德什 - 雷尼随机网络模型中，巨连通分支的相对大小关于平均度$\langle k \rangle$的函数。从图中可以看出，该函数在$\langle k \rangle = 1$处有一个相变，该相变标志着具有非零相对大小的巨连通分支出现了。

（b）～（e）示例随机网络及其在4种状态下的性质。

亚临界状态：$0 < \langle k \rangle < 1$（$p < \frac{1}{N}$，图 2-7b）

当$\langle k \rangle = 0$时，网络由N个孤立节点构成。$\langle k \rangle$的增加意味着我们往网络中增加了$N\langle k \rangle = pN(N-1)/2$个链接。然而，由于$\langle k \rangle < 1$，网络中只有少量的链接。因此在该状态下，网络中只有一些较小的连通分支（图2-7b）。

当然，任何时候我们都可以将最大连通分支指定为巨连通分支。然而，在这种状态下，最大连通分支的相对大小N_G/N接近于0。原因是：当$\langle k \rangle < 1$时，最大连通分支是一个大小为$N_G \sim \ln N$的树，最大连通分支大小的增加速度比网络大小的增加速度要慢得多。因此，在$N \to \infty$的极限情况下，$N_G/N \backsimeq \ln N/N \to 0$。

总之，亚临界状态下，网络由许多较小的连通分支组成，这些分支的大小服从指数分布（公式2.35）。因此，这些连通分支的大小相差不大，没有哪个连通分支的大小明显高于其他连通分支从而可以被指定为巨连通分支。

临界点：$\langle k \rangle = 1$（$p = \frac{1}{N}$，见图 2-7c）

临界点是网络从没有巨连通分支（$\langle k \rangle < 1$）变化到有巨连通分支（$\langle k \rangle > 1$）的边界。当网络处于临界点时，最大连通分支的相对大小仍然趋近于零（图2-7c）。实际上，此时最大连通分支的大小为$N_G \sim N^{2/3}$。因此，最大连通分支的大小N_G比网络大小增长得慢得多，在$N \to \infty$的极限情况下，最大连通分支的相对大小$N_G/N \sim N^{-1/3} \to 0$。

不过，需要注意的是，最大连通分支的绝对大小在$\langle k \rangle = 1$处有一个明显的跳跃。例如，对于有$N = 7 \times 10^9$个节点的随机网络——大小和全球社交网络相当，当$\langle k \rangle < 1$时，最大连通分支的大小为$N_G \backsimeq \ln N = \ln(7 \times 10^9) \backsimeq 22.7$。相比之下，当$\langle k \rangle = 1$时，$N_G \sim N^{2/3} = (7 \times 10^9)^{2/3} \backsimeq 3 \times 10^6$，比$\langle k \rangle < 1$时提升了约5个数量级。然而，无论处于亚临界状态还是临界点，最大连通分支都只包含了网络所有节点中很小的一部分。

总之，在临界点，大多数节点出现在数量众多的小连通分支中，这些连通分支大小的分布服从公式2.36。幂律形式意味着，大小差异很大的连通分支并存。为数众多的小连通分支主要是树，而最大连通分支则可能包含环。注意，处于临界点状态的网络，其很多性质和处于相变状态的物理系统的性质相似（进阶阅读2.F）。

超临界状态：$\langle k \rangle > 1$（$p > \frac{1}{N}$，图 2-7d）

这个状态最接近于真实系统，该状态下的最大连通分支开始像网络了。在临界点附近，最大连通分支大小的变化遵循：

$$N_G / N \sim \langle k \rangle - 1 \qquad (2.12)$$

或者：

$$N_G \sim (p - p_c) N \qquad (2.13)$$

这里的p_c由公式2.11得到。换句话说，巨连通分支的相对大小

不再是零。距离临界点越远，属于巨连通分支的节点所占的比例越大。注意，公式**2.12** 只在$\langle k \rangle = 1$附近有效。对于大的$\langle k \rangle$，N_G 和$\langle k \rangle$之间的依赖关系是非线性的（图**2-7a** ）。

总之，在超临界状态下，许多孤立的小连通分支和一个巨连通分支并存，这些连通分支的大小服从分布（公式**2.35** ）。这些小连通分支大多是树，而巨连通分支则由环或回路构成。超临界状态一直持续到所有节点都被巨连通分支吸收为止。

全连通状态：$\langle k \rangle > \ln N$（$p > \frac{\ln N}{N}$，**图 2-7e**）

当p 足够大时，巨连通分支吸收了所有的节点，此时有$N_G \simeq N$。由于没有孤立节点存在，整个网络变成连通的了。这种状态发生时，网络的平均度取决于N（进阶阅读**2.E** ）：

$$\langle k \rangle = \ln N \qquad (2.14)$$

注意，在刚进入连通状态时，网络仍然是相对稀疏的。原因是：对于大的N，$\ln N / N \to 0$。只有当$\langle k \rangle = N-1$时，网络才变成完全图。

总之，随机网络模型预测，网络的出现不是一个流畅的渐进过程：$\langle k \rangle$ 较小时观测到的孤立节点和小连通分支会经历一个相变，坍缩成一个巨连通分支（进阶阅读**2.F** ）。通过改变$\langle k \rangle$，我们可以观测到4个拓扑结构不同的状态（图**2-7** ）。

上述探讨采用的是一种实证的视角，该视角适用于将随机网络和真实系统进行对比。另外一种有着丰富内涵的视角来自数学文献（边栏**2.5** ）。

边栏**2.5**

图论中的网络演化

在随机图论的文献中，人们往往假设链接概率$p(N)$正比于N^z，这里的z 是介于$-\infty$和0之间的一个可调参数[15] 。埃尔德什和雷尼

发现，调节z 的值后，随机图的一些关键性质会突然出现。

如果一个图具有某种性质Q 的概率在$N \to \infty$时接近于1，我们会说该图具有性质Q 。也就是说，给定z 的情况下，要么几乎所有的图都具有性质Q ，要么几乎没有图具有性质Q 。例如，当z 小于-3/2时，几乎所有的图都只包含孤立节点和成对相连的点对。一旦z 超过-3/2，大多数网络都含有连接3个或更多节点的路径（图**2-8** ）。

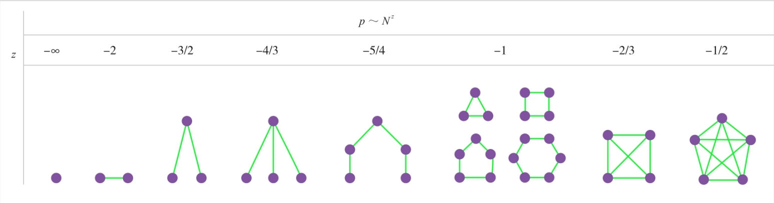


图**2-8**　随机图的演化

在链接概率p （N ）$\sim N^z$ 的随机图中，不同子图在随机图中出现的阈值概率。这里，阈值概率通过幂指数z 来刻画。当z ＜-3/2时，图中只有孤立节点和边；当z 的取值超过-3/2时，图中开始出现3阶树；当z ＜-4/3时，4阶树出现了；当z ＝-1时，各阶的树都出现了，同时出现的还有各阶的环；4阶完全图在z ＝-2/3时出现，并且随着z 的取值进一步增加，更高阶数的完全图开始出现。[19]

2.7
真实网络是超临界的

随机网络理论的两个预测对真实网络有着直接的重要性：

（**1**） 一旦平均度超过$\langle k \rangle$ ＝1，巨连通分支将会出现。因此，只有在$\langle k \rangle > 1$时，节点才能自组织成一个网络。

（**2**） 当$\langle k \rangle > \ln N$ 时，所有连通分支被巨连通分支吸收，从而形成一个连通网络。

真实网络满足巨连通分支存在的条件（$\langle k \rangle > 1$ ）吗？对于$\langle k \rangle > \ln N$ 的真实网络，巨连通分支是否包含所有节点，还是说巨连通分支之外还存在一些孤立节点或小连通分支？为了回答这些问题，我们

在给定$\langle k\rangle$的情况下将真实网络的结构和随机网络的理论预测进行对比。

测量表明，真实网络的平均度大大超过了临界阈值$\langle k\rangle=1$。实际上，据社会学家估计，一个普通人大约有1 000个相识的人；人脑中一个典型的神经元有大约7 000个突触；细胞中每个分子参与多个化学反应。

表2-1 中列出了几个无向网络的平均度，每个网络的平均度都满足$\langle k\rangle>1$。实际上，真实网络的平均度远远超出了临界阈值$\langle k\rangle=1$，这意味着真实网络都包含巨连通分支。**表2-1** 中列出的参考网络都如此。

表2-1 真实网络是连通的吗？

网络	N	L	$\langle k\rangle$	$\ln N$
互联网	192 244	609 066	6.34	12.17
电网	4 941	6 594	2.67	8.51
科学合作网络	23 133	94 439	8.08	10.05
演员网络	702 388	29 397 908	83.71	13.46
蛋白质相互作用网络	2 018	2 930	2.90	7.61

表中列出了几个无向网络的节点数N和链接数L，同时列出的网络性质还有$\langle k\rangle$和$\ln N$。根据随机网络理论，当$\langle k\rangle>1$时，网络中存在巨连通分支，而当$\langle k\rangle>\ln N$时，网络中的所有节点都属于巨连通分支。表中列出的所有网络都满足$\langle k\rangle>1$，但是大多数网络的平均度$\langle k\rangle$都在阈值$\ln N$之下（图**2-9**）。

我们现在考虑随机网络理论的第二个预测，看一下真实网络是仅由一个巨连通分支构成（$\langle k\rangle>\ln N$），还是除巨连通分支之外还包含多个连通分支（$\langle k\rangle<\ln N$）。对于全球的社会网络，从超临界状态到连通状态的相变点发生在$\langle k\rangle>\ln(7\times10^9)\approx22.7$。也就是说，如果一个普通人有超过23个相识的人，随机社会则仅由一个巨连通分支构成，没有孤立的人存在。由于$\langle k\rangle\approx1\ 000$，上述条件明显是满足的。不过，**表2-1** 中的许多真实网络并不满足网络处于连通状态的条件。因此，根据随机网络理论，这些网络会分割成多个孤立的连通分支。对于互联网，这是一个令人不安的预测，因为这意味着一些路由器和巨连通分支是不连通的，因而无法和其他路由器进行通信。对于电网而言，这一预测同样会带来问题，因为这意味着有些用户将无法使用电力。很明显，这些预测是不符合现实的。

总之，我们发现大多数真实网络处于超临界状态（图2-9　）。因此，这些网络中存在巨连通分支，这与观察结果是一致的。不过，除了少数几个真实网络外，巨连通分支和一些彼此不连通的小连通分支同时存在。注意，这些预测只有当真实网络可以由埃尔德什 - 雷尼模型准确描述（真实网络是随机的）时才是有效的。在接下来的章节中，随着对真实网络的结构了解得更多，我们就会明白为什么真实网络在不满足$\langle k\rangle > \ln N$ 的条件下仍然可以保持连通。

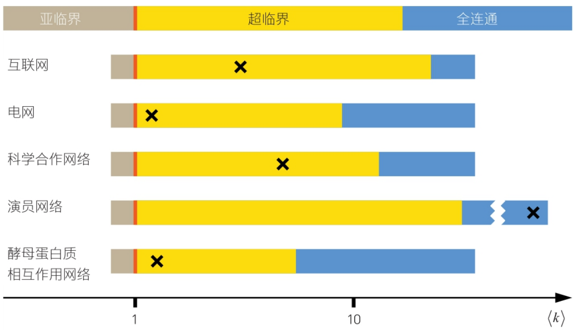


图2-9　大部分真实网络是超临界的

随机网络理论预测了网络的4种状态，图中用叉号标出了表2-1　中无向网络的平均度$\langle k\rangle$ 所处的位置。可以看出，大多数网络处于超临界状态，也就是说这些网络中包含了多个不相连的连通分支。只有演员网络处于全连通状态，即所有节点都是巨连通分支的一部分。注意，对于任意网络，亚临界状态和超临界状态的边界都是$\langle k\rangle =1$，而超临界状态和全连通状态之间的边界$\langle k\rangle = \ln N$ 却是因网络而异的。

2.8
小世界

小世界现象——又称六度分隔，长期以来备受公众关注。小世界现象指出，在地球上任意选择两个人，你会发现他们之间相隔最多6个相识关系（图2-10　）。如果说生活在同一个城市里的两个人彼此之间只相隔少数几个相识关系，你可能不会感到吃惊。然而，小世界概念告诉我们，即便是生活在地球上相对方位的两个人，彼此之间也只相隔少数几个相识关系。

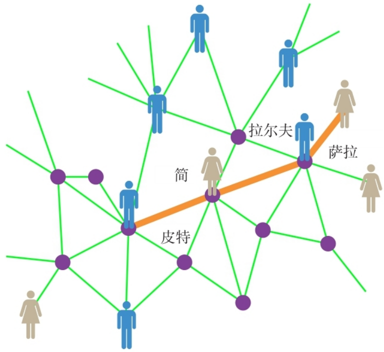


图2-10　六度分隔

根据六度分隔理论，来自世界上任何地方的两个人，都可以通过不超过6个相识关系连接起来。例如，虽然萨拉并不认识皮特，但她认识拉尔夫，而拉尔夫认识简，简又认识皮特，因此萨拉和皮特之间相隔3个相识关系，或者说萨拉和皮特是三度分隔的。在网络科学的语言中，六度分隔被称为小世界性质——意味着网络中任意两个节点之间的距离非常小。

在网络科学的语言中，小世界现象意味着，网络中随机选择的两个节点之间的距离很短。这一论断带来了两个问题：怎样才算短，或者说和什么相比是短的呢？如何解释距离短这一现象呢？

这两个问题都可以通过一个简单的计算来回答。考虑一个平均度为$\langle k\rangle$的随机网络，平均情况下每个节点有：

$\langle k\rangle$个距离为1的节点（$d=1$）；

$\langle k\rangle^2$个距离为2的节点（$d=2$）；

$\langle k\rangle^3$个距离为3的节点（$d=3$）；

……

$\langle k\rangle^d$个距离为d的节点。

例如，如果$\langle k\rangle\approx1\ 000$（每个人平均认识的人数），那么从某个人出发距离为2的人数约为10^6，距离为3的人数大约有10亿人——和全球总人数在一个数量级。

精确地讲，从一个起始节点出发距离不超过d的期望节点数为：

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \cdots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1} \qquad (2.15)$$

注意，$N(d)$不能超过网络中的节点总数N。因此，距离不可能取任意大的值。最大距离d_{max}——又称网络直径，可以通过下面的公式算出：

$$N(d_{max}) \approx N \qquad (2.16)$$

假设$\langle k \rangle \gg 1$，公式**2.15**的分子和分母中的（-1）项可以忽略不计，从而有：

$$\langle k \rangle^{d_{max}} \approx N \qquad (2.17)$$

因此，随机网络的直径为：

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle} \qquad (2.18)$$

公式**2.18**是小世界现象的数学表示。关键是对公式**2.18**的解读：

- 公式**2.18**给出了网络直径d_{max}和网络大小N之间的关系。然而，对大多数网络而言，和对网络直径d_{max}的近似相比，公式**2.18**更好地提供了两个随机选择节点之间平均距离$\langle d \rangle$的近似（表**2-2**）。这是因为，d_{max}往往取决于少数几个极端路径；$\langle d \rangle$是关于所有节点对的平均值，受随机波动影响不大。因此，通常情况下，小世界性质定义为：

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle} \qquad (2.19)$$

表**2-2**　六度分隔

网络	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln\langle k \rangle}$
互联网	192 244	609 066	6.34	6.98	26	6.58
万维网	325 729	1 497 134	4.60	11.27	93	8.31
电网	4941	6594	2.67	18.99	46	8.66
手机通话网络	36 595	103 731	1.81	5.88	18	18.4
电子邮件网络	57 194	103 731	1.81	5.88	18	18.4
科学合作网络	23 133	93 439	8.08	5.35	15	4.81
演员网络	702 388	29 397 908	83.71	3.91	14	3.04
引文网络	449 673	4 707 958	10.43	11.21	42	5.55
代谢网络	1 039	5 802	5.58	2.98	8	4.04
蛋白质相互作用网络	2 018	2 930	2.90	5.61	14	7.14

表中给出了10个网络的平均距离$\langle d \rangle$ 和最大距离d_{max} 。最后一列是公式**2.19** 预测的$\langle d \rangle$ ，可以看出它是对实测$\langle d \rangle$ 的一个合理近似。不过，公式**2.19** 的预测并不完美：在下一章我们将看到，对于很多真实网络，公式**2.19** 需要调整。对于有向网络，平均度和路径长度是考虑了边的方向测量出的。

公式**2.19** 刻画了平均距离和网络大小N 、平均度$\langle k \rangle$ 之间的依赖关系。

- 一般而言，$\ln N \ll N$ 。因此，平均距离$\langle d \rangle$ 对$\ln N$ 的依赖，意味着随机网络中的距离比网络大小要小多个数量级。因此，“小世界现象”中的“小”是指，平均路径长度或网络直径和网络大小的关系是对数关系。因此，“小”的意思是，$\langle d \rangle$ 正比于$\ln N$ ，而不是正比于N 或者N 的幂（图**2-11** ）。

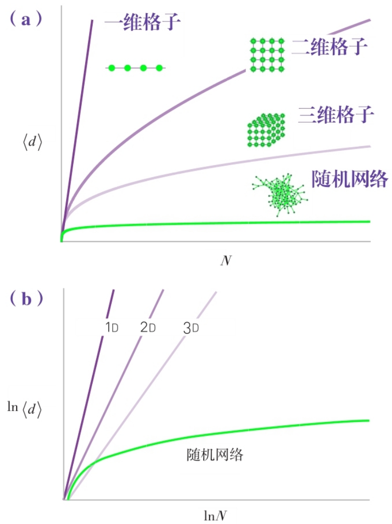


图2-11　为什么小世界令人惊奇？

我们关于距离的直觉大多来自认识规则格子网络中的经验，而格子网络不具有小世界性质：

一维（1D）：对于一维格子（一条长度为N的线），其直径及平均路径长度都和N呈线性关系：$d_{max} \sim \langle d \rangle \sim N$。

二维（2D）：对于正方形格子，有$d_{max} \sim \langle d \rangle \sim N^{1/2}$。

三维（3D）：对于立方体格子，有$d_{max} \sim \langle d \rangle \sim N^{1/3}$。

四维（4D）：通常情况下，对于d维格子，有$d_{max} \sim \langle d \rangle \sim N^{1/d}$。

上述多项式依赖关系表明，格子网络的网络直径或平均路径长度随N增长的速度要比公式**2.19**快得多。这意味着，格子网络中的平均路径长度明显大于随机网络中的平均路径长度。例如，如果社会网络是一个正方形格子（2D），每个人只认识其邻居，那么任意两个人之间的平均距离将大致为$(7\times10^9)^{1/2}=83\ 666$。即使我们把每个人平均认识的人数从二维格子的4修正为1 000，人们之间的平均距离也比公式**2.19**要高几个数量级。

（**a**）线性坐标下，格子网络和随机网络中$\langle d \rangle$对N的依赖关系。

（**b**）和（a）相同，不过坐标换成了双对数坐标。

- 公式**2.19**中的项$1/\ln\langle k \rangle$意味着，网络越稠密，节点间的距离越小。
- 真实网络中，公式**2.19**需要进行系统修正。这源于这样

一个事实：当$d > \langle d \rangle$时，和起始节点距离为d的节点数目迅速减少（进阶阅读**2.F**）。

我们现在来看一下公式**2.19**对社交网络意味着什么。全球社交网络中，节点数为$N \approx 7 \times 10^9$，平均度为$\langle k \rangle \approx 10^3$，因此有：

$$\langle d \rangle \approx \frac{\ln 7 \times 10^9}{\ln 10^3} = 3.28 \qquad (2.20)$$

因此，地球上所有人只相隔3~4个相识关系[20]。公式**2.20**的估计可能比人们经常提到的“六度”更接近真实值（边栏**2.7**）。

边栏**2.7**

六度：实验验证

关于小世界现象的第一个实证研究发生在1967年。斯坦利·米尔格拉姆在波普尔和科亨[20]的工作基础上设计了一个实验，来测量社会网络中的距离[24]，[25]。米尔格拉姆选择了波士顿的一位股票经纪人和马萨诸塞州沙伦神学院的一位学生作为目标，然后在威奇托和奥马哈随机选择一些居民参与实验，给他们每人发送一封信，信的内容包括对此项实验研究目的简要介绍，以及目标对象的照片、姓名、地址和其他相关信息。实验要求每位参与者从他的朋友、亲戚或熟人中选出他认为最有可能认识目标对象的人，然后将信转给他。

几天后，第一封信到达了，只经过了两个链接。最终，全部296封信中有64封回来了。不过，有些信则经过了多达10余个链接才送到目标手中[25]。使用这些成功返回的信件，米尔格拉姆得以算出将信从原发者送达目标对象中间需要经过的人数（图**2-12a**）。他发现，中间所经历人数的平均值为5.2，这个相对而言很小的数字，非常接近弗里杰什·考林蒂（Frigyes Karinthy）于1929年做出的判断（边栏**2.8**）。

米尔格拉姆没有人们之间的相识关系网络，因此他的实验无法

得到实验参与者和目标对象之间真实距离。今天，Facebook拥有了有史以来世界上最大的社交网络地图。使用Facebook 2011年5月的社交网络——包括7.21亿个活跃用户和680亿个对称朋友关系，研究人员发现这些用户之间的平均距离是4.74（图**2-12b**）。因此，研究发现的“四度分隔”[18]，比米尔格拉姆的六度[24]，[25]更接近公式**2.20**的预测。

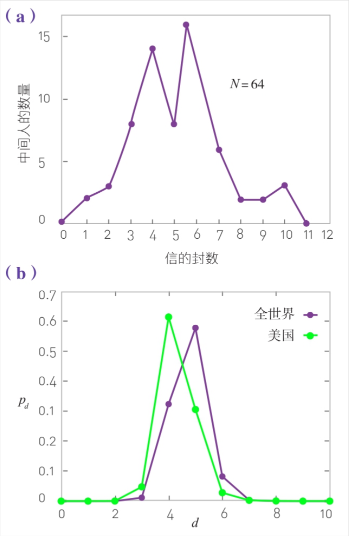


图**2-12**　六度？从米尔格拉姆到**Facebook**

（**a**）　在米尔格拉姆的实验中，296封信中的64封到达了收件人手中。图中给出了这64封邮件经过的中间人数的分布。从中可以看出，有些信只经过了一个中间人，有些信则经过了多达10个中间人。该分布的平均值为5.2，这表明平均需要约6次“联络”才能使信到达收件人那里。20年后，剧作家约翰·格尔将这一现象称为“六度分隔”[25]。

（**b**）　Facebook所有用户间的平均距离分布以及美国Facebook用户间的平均距离分布。基于Facebook社交网络的节点数N和链接数L，公式**2.19**预测，社交网络中人们的平均距离约为3.9，和文献[18]中报道的四度相差不大。

我问一位智者，按照实验中的方式从内布拉斯加州到沙伦，需要经历多少个中间人，他回答说，将需要不少于100个人。

——斯坦利·米尔格拉姆，1969

关于随机网络的小世界性质，包括公式**2.19**的结论，均出自由曼弗雷德·科亨和伊锡尔·德索拉·普尔合著的一篇并不怎么出名的论文。论文中给出了小世界性质的数学建模，并深入讨论了小世界性质的社会学意义。这篇论文启发了著名的米尔格拉姆实验（边栏**2.6**），而米尔格拉姆实验又带来了“六度分隔”的说法。

边栏**2.6**

十九度分隔

万维网中，需要多少次点击才能从一个随机选择的网页到达另外一个随机选择的网页呢？回答该问题的困难之处在于，我们没有完整的万维网地图，只有这个地图的采样。不过，我们可以不断增加万维网采样的大小并测量其平均路径长度，从而对整个万维网的平均路径长度进行估计，这一过程被称为有限标度法。测量结果表明，万维网的平均路径长度和万维网大小之间的关系为[21]：

$$\langle d\rangle \approx 0.35 + 0.89\ln N$$

1999年时，万维网大约有8亿个网页[22]，根据上述公式可以计算出当时万维网的平均路径长度为$\langle d\rangle \approx 18.69$。换句话说，1999年时两个随机选择的页面之间平均相隔19次点击，这被称为十九度分隔。随后，在一个有2亿个网页的万维网采样上得出的测量结果是$\langle d\rangle \approx 16$[23]，和公式预测的$\langle d\rangle \approx 17$很接近。目前，万维网估计有大约$N \sim 10^{12}$个网页，相应地有$\langle d\rangle \approx 25$。因此，$\langle d\rangle$不是固定不变的，而是随着网络大小的增加而增加。万维网的这个平均路径长度25比社会网络中众所周知的六度要大得多（边栏**2.7**）。这个差距很容易理解：万维网的平均度比社会网络要小，而大小却比社会网络大，根据公式**2.19**，这两方面使得万维网的直径要更大些。

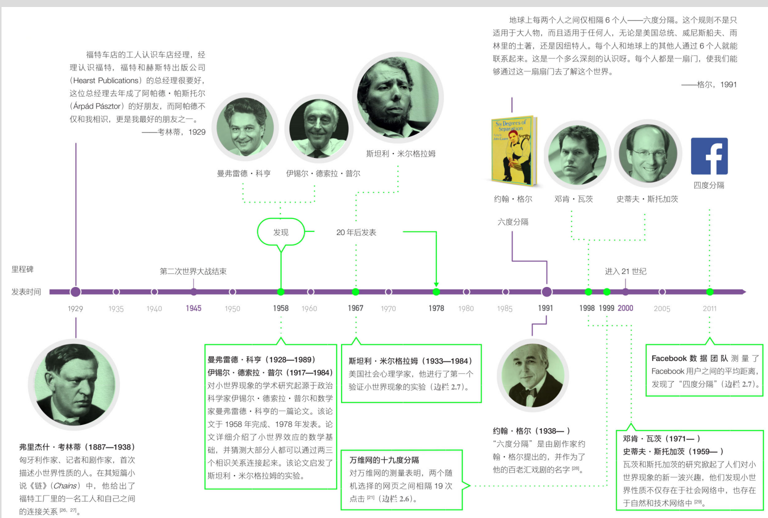
虽然小世界性质是在社会系统中被发现的，但它不只适用于社交网络（边栏**2.6**）。为了阐明这一点，我们在表**2-2**中将几个真实网

络的平均路径长度$\langle d\rangle$ 和公式**2.19** 的预测值进行了对比。对比发现，尽管这些网络形态多样、大小N 以及平均度$\langle k\rangle$ 相差很大，公式**2.19**也给实际观测到的$\langle d\rangle$ 提供了一个很好的近似。

总之，小世界性质不仅引发了公众的想象（边栏**2.8** ），在网络科学中也起到了重要作用。小世界现象在随机网络模型背景下可以得到非常好的解释：小世界现象源于这样一个事实：从某个节点出发，距离为d 的节点个数随着d 指数增长。我们在接下来的章节中将看到，真实网络和公式**2.19** 之间存在系统性偏差，这迫使我们用更准确的预测取代公式**2.19** 。不过，随机网络模型提供的关于小世界现象起源的直观认识在真实网络中仍然是有效的。

边栏**2.8**

万维网的十九度分隔



2.9 集聚系数

一个节点的度未包含该节点邻居之间关系的信息。这些邻居节点是彼此相连的还是相互孤立的？这个问题可以用局部集聚系数C_i来回答，C_i反映了节点i的直接邻居之间的链接密度：$C_i=0$意味着节点i的邻居之间没有链接，$C_i=1$则意味着节点i的任意两个邻居之间都相互连接（见**1.10**节）。

对于随机网络中的一个节点i，要计算其集聚系数C_i，我们需要估计出该节点的k_i个邻居之间的链接数L_i。随机网络中，节点i的两个邻居之间的链接概率是p。由于节点i的k_i个邻居之间最多有$k_i(k_i-1)/2$条链接，L_i的期望值为：

$$\langle L_i \rangle = p\frac{k_i(k_i-1)}{2} \quad (2.20)$$

因此，随机网络的局部集聚系数为：

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i-1)} = p = \frac{\langle k \rangle}{N} \quad (2.21)$$

公式**2.21**给出了两个预测：

（**1**）给定$\langle k \rangle$的情况下，网络越大，节点的集聚系数越小。因此，节点的局部集聚系数和$1/N$成正比。注意，网络的平均集聚系数$\langle C \rangle$也服从公式**2.21**。

（**2**）随机网络中，节点的局部集聚系数和节点的度相互独立。

为了测试公式**2.21**的有效性，我们绘制了几个无向网络的$\langle C \rangle / \langle k \rangle$关于网络大小$N$的函数（图**2-13a**）。我们发现，$\langle C \rangle / \langle k \rangle$并非按照$N^{-1}$的速度减少，而是与$N$大体上相互独立，这一点和公式**2.21**的第一个预测不符。图**2-13b**～**d**展示了三个真实网络中C对节点度k_i的依赖性，我们发现$C(k)$随着节点度增加而下降，违背了公式**2.21**的第二个预测。

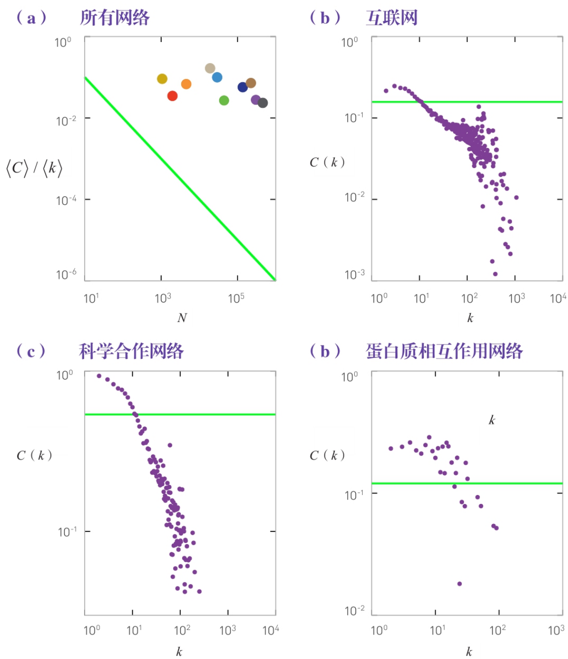


图2-13　真实网络的集聚系数

（a） 真实网络的集聚系数和公式2.21 预测的随机网络的集聚系数。圆圈及其颜色对应表2-2 中的网络。为了计算$\langle C\rangle$ 和$\langle k\rangle$ ，我们将有向网络视为无向网络。绿色的线对应公式2.21 预测的随机网络的集聚系数——按照N^{-1} 减少。与之相反，真实网络中，集聚系数$\langle C\rangle$ 似乎与N 是相互独立的。

（b） ～（d） 局部集聚系数$C(k)$对（b）互联网、（c）科学合作网络、（d）蛋白质相互作用网络节点度的依赖性。$C(k)$是节点度为k 的所有节点的平均局部集聚系数。绿色水平线对应的是$\langle C\rangle$ 。

总之，随机网络模型不能刻画真实网络的集聚特性。相反，在具有同样的N 和L 的情况下，真实网络的集聚系数比随机网络预测的集聚系数要高得多。瓦茨和斯托加茨对随机网络模型进行了扩展 [29] ，解决了高集聚系数$\langle C\rangle$ 和小世界性质共存的问题（边栏2.9 ）。然而，扩展后的随机网络模型仍然不能解释为什么大度节点的集聚系数比小度节点的集聚系数要小。第8章 将探讨能够解释$C(k)$形状的模型。

边栏2.9

瓦茨 - 斯托加茨模型

邓肯·瓦茨和史蒂夫·斯托加茨对随机网络模型进行了扩展（图**2-14**），该扩展基于如下两个观察[29]：

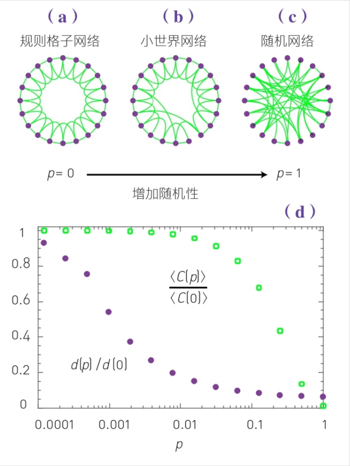


图**2-14**　瓦茨 - 斯托加茨模型[29]

（**a**）　首先从节点环开始，每个节点和它的直接邻居及二阶邻居相连。因此，初始时，每个节点有$\langle C\rangle=3/4$（$p=0$）。

（**b**）　每条链接以概率p重连到一个随机选择的节点。当p较小时，网络仍然保持高聚集性，不过随机连接带来的“长程边”可以大大缩短节点间的距离。

（**c**）　当$p=1$时，所有链接都进行了重连，网络变成了随机网络。

（**d**）　平均路径长度$d(p)$和集聚系数$\langle C(p)\rangle$对重连参数p的依赖性。注意，$d(p)$和$\langle C(p)\rangle$已经分别使用$d(0)$和$\langle C(0)\rangle$进行了归一化，这里的$d(0)$和$\langle C(0)\rangle$是从最初的规则格子网络中得出的（图a所示的$p=0$的情况）。$d(p)$的快速下降标志着小世界现象的发生。在$d(p)$的下降过程中，$\langle C(p)\rangle$仍然保持很高的数值。因此，在$0.001<p<0.1$的范围内，短网络路径长度和高聚集性共存。这里的所有图，都有$N=1\ 000$和$\langle k\rangle=10$。

（1）小世界性质

真实网络中，两个节点之间的平均距离对网络大小N的依赖关系是对数形式（公式**2.18**），而不是规则格子网络中的多项式形式（图**2-11**）。

（2）高集聚系数

真实网络的平均集聚系数远远高于具有类似N 和L 的随机网络（图**2-13a** ）。

瓦茨 - 斯托加茨模型（也称小世界模型）介于规则格子网络和随机网络之间：规则格子网络具有高集聚系数，但没有小世界性质；随机网络具有小世界性质，但集聚系数低（图**2-14a** ~**c** ）。数值模拟表明，在链接重连参数的很大取值范围内，网络同时具有低平均路径长度和高集聚系数，从而重现了高集聚系数和小世界性质共存的现象（图**2-14d** ）。

作为随机网络模型的一种扩展，瓦茨 - 斯托加茨模型预测，节点的度分布类似泊松分布。因此，像图**2-6** 中展示的那样，网络中不存在大度节点。此外，瓦茨 - 斯托加茨模型预测，$C(k)$和k 无关，无法刻画图**2-13b** ~**d** 中观察到的$C(k)$和k 之间的相关性。我们在下一章中将会看到，要想理解小世界性质和高集聚系数共存的现象，必须首先能够正确刻画网络的度分布。

2.10 小结：真实网络不是随机的

自1959年被提出以来，随机网络模型就成了研究复杂网络的主流数学方法。该模型认为，复杂系统中观察到的看上去随机的网络应该被描述成纯粹随机的网络。如此一来，随机网络模型将复杂性等同于随机性。因此，我们不禁要问：

我们真的相信真实网络是随机的吗？

答案显然是否定的。例如，蛋白质间的相互作用受严格的生物化学定律支配，因此，细胞要发挥其功用则其化学结构不可能是随机的。同样，在随机社会中，对一个美国学生而言，他的某个同学和中国工厂的某个工人成为他朋友的概率几乎相当。这似乎不合乎逻辑。

现实中，我们猜测大部分复杂系统的背后有一个深层次的秩序存

在。这一秩序必然会反映在描述复杂系统架构的网络结构中，从而导致其网络结构和纯粹的随机网络之间存在系统性偏差。

随机网络在多大程度上可以刻画真实系统，不能由认识论来决定，而应该通过系统的定量比较来决定。如前文所述，随机网络理论做出了许多定量预测，这使得我们可以定量地判定随机网络在多大程度上可以刻画真实系统：

度分布

随机网络的度分布是一个二项分布，在$k \ll N$的极限情况下可以通过泊松分布进行很好的近似。然而，如图**2-5**所示，泊松分布不能刻画真实网络的度分布。真实系统中，大度节点的数量要比随机网络模型预测的多得多。

连通性

随机网络理论预测，在$\langle k \rangle > 1$时，网络中会出现巨连通分支。本章探讨过的所有网络都满足这一条件。不过，这些网络大多不满足$\langle k \rangle > \ln N$的条件（表**2-1**）。这意味着，除了巨连通分支，这些网络还包含一些孤立的小连通分支。实际上，除去少数几个网络，大部分真实网络中并不包含孤立的小连通分支。

平均路径长度

随机网络理论预测，平均路径长度服从公式**2.19**。这是对实际观测路径长度的一个合理近似。因此，随机网络模型可以解释小世界现象的出现。

集聚系数

随机网络中，局部集聚系数和节点的度无关，$\langle C \rangle$以$1/N$依赖于网络大小N。相比之下，真实网络中实际测量出的$C(k)$随着节点度的增大而减小，且与网络大小基本无关（图**2-13**）。

总之，小世界现象是唯一可以由随机网络模型合理解释的性质。在真实网络中，从度分布到集聚系数等所有其他网络特性，都与随机网络有着显著差异。瓦茨和斯托加茨对埃尔德什 - 雷尼模型的扩展，成功地解释了高集聚系数C 和低平均路径长度$\langle d \rangle$ 的共存，却不能解释度分布和$C(k)$。事实上，我们接触的真实网络越多，就越容易得出如下结论：没有任何一个真实网络可以由随机网络模型准确刻画。

边栏2.11

随机网络一览

定义：N 个节点，每对节点之间的连接概率是p 。

平均度：

$$\langle k \rangle = p(N-1)$$

平均链接数：

$$\langle L \rangle = \frac{pN(N-1)}{2}$$

度分布：

二项形式：

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

泊松形式：

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

巨连通分支（N_G ）：

$$\langle k \rangle < 1 : N_G \sim \ln N$$

$$1 < \langle k \rangle < \ln N : N_G \sim N^{\frac{2}{3}}$$

$$\langle k \rangle > \ln N : N_G \sim (p - p_c) N$$

平均距离：

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$$

集聚系数：

$$\langle C \rangle = \frac{\langle k \rangle}{N}$$

这一结论自然而然地引出了一个疑问：既然真实网络不是随机的，为什么我们还要用一个章节来介绍随机网络模型呢？答案很简单：在继续探讨真实网络的性质时，随机网络模型将会为我们提供一个重要参照。每当观测到某种网络性质时，我们都会问，该性质是否只是偶然出现的。为此，我们可以使用随机网络模型作为指南：如果该性质在随机网络中存在，则意味着它可以由随机性解释。如果该性质在随机网络中不存在，则它很可能标志着某种秩序——需要更深入的解释。因此，尽管随机网络模型对大多数真实系统而言可能是错误的，它对于网络科学仍然十分重要（边栏**2.10**）。

边栏**2.10**

随机网络和网络科学

随机网络和真实网络之间的不吻合提出了一个重要的问题：既然随机网络理论和真实网络如此不吻合，那么该理论为什么能够存活这么久呢？答案很简单：随机网络理论从来没有打算作为真实系统的模型。

埃尔德什和雷尼在他们的第一篇论文 [2] 中写道，随机网络“不仅从纯数学角度来看是有趣的。实际上，图的演化也可以视为某个国家或单位的通信网（例如铁路网、公路网或电网等）演化的简化

模型”。然而，在他们关于随机网络的8篇系列论文 [2-9] 中，这是唯一一次提到随机网络理论潜在应用价值的地方。随机网络理论随后的发展则完全是由该问题的内在数学挑战驱动，而不是由其应用价值驱动的。

按照托马斯·库恩（Thomas Kuhn）的范式理论，人们很容易将网络科学视为从随机图论到真实网络理论的范式变化 [30] 。实际上，20世纪90年代末以前，网络范式还不存在。在那个时期，人们还没有系统地将图论模型和真实网络的性质进行对比。埃尔德什和雷尼的工作直到网络科学出现之后才取得数学领域之外的影响（图**2-15**）。

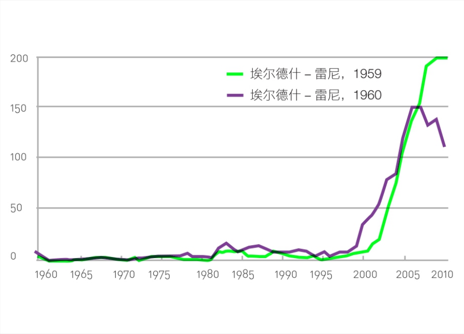


图**2-15**　网络科学和随机网络

今天我们将埃尔德什 - 雷尼模型视为网络理论的基础，而在此之前该模型的影响仅局限于数学的一个小分支。这一点可以从埃尔德什和雷尼在1959年和1960年发表前两篇关于随机网络的论文 [2] ，[3] 的年引用数量中清楚地看出来。在发表后的40年间，这两篇论文每年获得的引用数不超过10。不过，在无标度网络的第一批论文 [21] ，[31] ，[32] 将埃尔德什和雷尼提出的随机网络理论作为网络科学的参照模型之后，这两篇论文的引用数有了爆炸式的增长。

网络科学不但没有降低埃尔德什和雷尼的贡献，反而获得了他们起初不曾想到的影响。我们讨论随机网络和真实网络之间的差异，主要出于教学原因：为我们随后理解真实网络的性质提供一个合适的基础。

2.11

课后习题

习题1　埃尔德什 - 雷尼网络

假如一个埃尔德什 - 雷尼网络的节点数为$N=3\ 000$，节点间的连接概率为$p=10^{-3}$。

（1）该网络的期望链接数$\langle L\rangle$是多少？

（2）该网络处于哪个状态？

（3）计算该网络处于临界点时的连接概率p_c。

（4）假如连接概率$p=10^{-3}$固定不变，计算出只有一个连通分支存在时的网络大小N^{cr}。

（5）对于（4）中的网络，计算出其平均度$\langle k^{\mathrm{cr}}\rangle$和节点平均距离$\langle d\rangle$。

（6）计算出该网络的度分布p_k（采用泊松分布近似）。

习题2　生成埃尔德什 - 雷尼网络

基于随机网络模型$G(N,p)$，使用计算机生成大小为$N=500$的三个网络，平均度分别为（1）$\langle k\rangle=0.8$，（2）$\langle k\rangle=1$和（3）$\langle k\rangle=8$，并对这三个网络进行可视化。

习题3　环形网络

考察大小为N的环形网络，每个节点和其每侧m个邻居节点相连（因此，每个节点的度为$2m$）。图**2-14a**给出了一个例子，该网络中$m=2$，$N=20$。计算该网络的平均集聚系数$\langle C\rangle$和平均最短路径长度$\langle d\rangle$。简单起见，我们假设N和m的选择满足$(n-1)/2m$是整数。当$N\gg 1$时，$\langle C\rangle$会怎样？$\langle d\rangle$会怎样？

习题4　凯莱树

凯莱树是从一个度为k的中心节点出发构造出来的一棵对称树。和中心节点相距为d（$d<P$）的每个节点的度为k。和中心节点相距为P的每个节点的度为1，这些节点被称为叶子节点（图**2-16**是一棵$k=3$、$P=5$的凯莱树）。

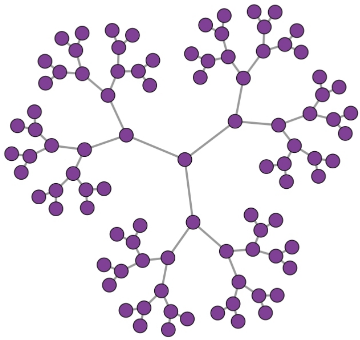


图**2-16**　凯莱树

一棵$k=3$、$P=5$的凯莱树。

（**1**）计算从中心节点出发t步可以到达的节点个数。

（**2**）计算网络的度分布。

（**3**）计算直径$d_{\max}$。

（**4**）写出直径$d_{\max}$关于节点总数N的表达式。

（**5**）该网络具有小世界性质吗？

习题5　“势利”网络

考虑一个有N个红色节点和N个蓝色节点的网络。相同颜色节点之间的连接概率为p，不同颜色节点之间的连接概率为q。如果$p>q$，则称该网络是“势利的”。“势利”是指节点倾向于和具有相同颜色的其他节点相连。当$q=0$时，网络中至少包含两个连通分支，分别对应两种颜色的节点。

（**1**）计算只有蓝色节点构成的“蓝色”子网络的平均度和整个网络的平均度。

（2） 给出网络大概率只包含一个连通分支所需要的p 和q 。

（3） 证明：当N 很大时，非常“势利”的网络（$p \gg q$ ）也具有小世界性质。

习题6 “势利”社交网络

考虑上面所讨论模型的如下变种：网络有$2N$ 个节点，红色节点和蓝色节点数据相同，$2N$ 个节点中有比例为f 的节点是紫色节点。蓝色节点和红色节点之间不相连（$q=0$），蓝色节点之间的连接概率是p ，红色节点之间的连接概率也是p 。另外，紫色节点和红色、蓝色节点之间的连接概率也是p 。

（1） 如果红色节点和蓝色节点之间的距离为2，我们则称红色社区和蓝色社区是彼此交互的。计算出此时所需要的紫色节点的比例。

（2） 如果蓝色（或红色）节点的度$\langle k \rangle \gg 1$，计算紫色社区中的节点数。

（3） 该模型对于研究社会网络结构的启发有哪些？

2.12 进阶阅读2.A 推导泊松分布

为了推导泊松形式的度分布，让我们从刻画随机网络度分布的精确二项分布（公式**2.7** ）开始：

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \quad (2.22)$$

我们将公式右侧的第一项重写为：

$$\binom{N-1}{k}=\frac{(N-1)(N-2)(N-3)\cdots(N-k)}{k!}\approx\frac{(N-1)^k}{k!} \quad (2.23)$$

这里，最后一项中我们使用了$k \ll N$。将公式**2.22** 的最后一项简化为：

$$\ln\left[(1-p)^{N-1-k}\right]=(N-1-k)\ln\left(1-\frac{\langle k\rangle}{N-1}\right)$$

进而，使用级数展开：

$$\ln(1+x)=\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots,\forall|x|\leqslant 1$$

从而，我们得到：

$$\ln\left[(1-p)^{N-1-k}\right]\approx(N-1-k)\frac{\langle k\rangle}{N-1}=-\langle k\rangle\left(1-\frac{k}{N-1}\right)\approx-\langle k\rangle$$

注意，该公式只在$N \gg k$ 时成立。这表明在公式推导过程中隐含着“小度近似”。因此，公式**2.22** 的最后一项变成：

$$(1-p)^{N-1-k}=e^{-\langle k\rangle} \quad (2.24)$$

结合公式**2.22** 、**2.23** 和**2.24** ，我们得到了度分布的泊松形式：

$$p_k=\binom{N-1}{k}p^k(1-p)^{N-1-k}=\frac{(N-1)^k}{k!}p^k e^{-k}$$

$$=\frac{(N-1)^k}{k!}\left(\frac{\langle k\rangle}{N-1}\right)^k e^{-\langle k\rangle}$$

或者写成

$$p_k=e^{-\langle k\rangle}\frac{\langle k\rangle^k}{k!} \quad (2.25)$$

2.13
进阶阅读2.B

最大度和最小度

为了确定随机网络中最大节点的期望度——网络的自然上界，我们定义度k_{max}，使得有N个节点的网络中最多只有一个节点的度比k_{max}大。数学上，这意味着泊松分布p_k对应于$k \geq k_{max}$区域的面积应趋近于0（图**2-17**）。该区域的面积大小可以使用度分布p_k的累积分布$P(k)$表示为$1-P(k_{max})$，从而网络中最大节点满足：

$$N\left[1-P\left(k_{max}\right)\right] \approx 1 \quad (2.26)$$

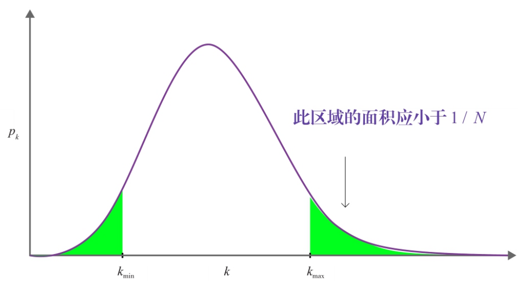


图**2-17**　最小度和最大度

网络的期望最大度k_{max}满足度大于k_{max}的节点数不超过1的条件，k_{max}通常被称为度分布的自然上界。为了计算k_{max}，我们需要让度分布p_k在区域$k>k_{max}$上的面积等于$1/N$，即处于该区域的节点数目正好为1。采用类似的方式，同样可以算出期望最小度k_{min}。

我们在这里使用≈代替＝，是因为k_{max}是一个整数，一般会使得精确情况下该方程无解。对于泊松分布，有：

$$1-P\left(k_{max}\right)=1-e^{-\langle k\rangle}\sum_{k=0}^{k_{max}}\frac{\langle k\rangle^k}{k!}=e^{-\langle k\rangle}\sum_{k=k_{max}+1}^{\infty}\frac{\langle k\rangle^k}{k!}\approx e^{-\langle k\rangle}\frac{\langle k\rangle^{k_{max}+1}}{\left(k_{max}+1\right)!} \quad (2.27)$$

在最后一项中，我们使用最大项来近似总和。

对于全球社会网络来说，其大致大小为$N=10^9$，平均度为$\langle k\rangle=1\ 000$。根据公式**2.26**和公式**2.27**，可以计算出$k_{max}=1\ 185$。这表明，随机社会网络中没有朋友数特别多的人，即枢纽节点。

采用类似的方法，我们可以计算出最小节点的期望度k_{min}。要求

最多只有一个节点的度小于$k_{\min}$，则：

$$NP\left(k_{\min}-1\right)\approx 1 \tag{2.28}$$

对于埃尔德什 - 雷尼网络，我们有

$$P\left(k_{\min}-1\right)=e^{-\langle k\rangle}\sum_{k=0}^{k_{\min}-1}\frac{\langle k\rangle^{k}}{k!} \tag{2.29}$$

公式**2.28** 在$N=10^9$、$\langle k\rangle=1\ 000$时的解为$k_{\min}=816$。

2.14
进阶阅读2.C
巨连通分支

本节将介绍“巨连通分支在$\langle k\rangle=1$时出现”[33] 这一论断。该论断由雷·索洛莫诺夫（Ray Solomonoff）和阿纳托尔·拉波波特（Anatol Rapoport）[11]、埃尔德什和雷尼[2] 独立提出。

我们用$u=1-N_G/N$ 表示不在巨连通分支（GC）中的节点比例，巨连通分支的大小记为N_G。如果节点i 是GC的一部分，则它必须和GC中的另外一个节点j 相连。因此，节点i 不是GC的一部分，可能有如下两个原因：

（**1**）i 和j 之间没有链接（这个概率是$1-p$）。

（**2**）i 和j 之间有链接，但j 不是GC的一部分（这个概率是pu）。

因此，节点i 通过j 和GC不相连的概率是$1-p+pu$。理论上讲，节点i 和可以通过除其自身之外的$N-1$个节点中的任意一个节点和GC相连，因此节点i 通过任何其他节点都和GC不相连的概率为$(1-p+pu)^{N-1}$。由于u 是不属于GC的节点比例，因此对于任何p 和N，巨连通分支的大小可以通过求解方程

$$u=\left(1-p+pu\right)^{N-1} \tag{2.30}$$

而得出，即$N_G = N(1-u)$。当$\langle k \rangle \ll N$时，利用$p = \langle k \rangle / (N-1)$并对方程两侧取对数，我们得出：

$$\begin{aligned} \ln u &= (N-1)\ln\left[1 - \frac{\langle k \rangle}{N-1}(1-u)\right] \\ &\approx (N-1)\left[-\frac{\langle k \rangle}{N-1}(1-u)\right] = -\langle k \rangle (1-u) \end{aligned} \qquad (2.31)$$

这里我们使用了$\ln(1+x)$的级数展开。

对公式**2.31** 的两边都取指数，得到$u = \exp[-\langle k \rangle(1-u)]$。我们使用$S = N_G / N$表示巨连通分支中的节点比例，则根据$S = 1-u$和公式**2.31** 有：

$$S = 1 - e^{-\langle k \rangle S} \qquad (2.32)$$

公式**2.32** 给出了巨连通分支的大小S关于平均度$\langle k \rangle$的函数（图**2-18**）。公式**2.32** 虽然看起来很简单，但没有闭式解。不过，我们可以通过图示的方式绘制出当$\langle k \rangle$取不同值时公式**2.32** 的右侧关于S的函数。公式**2.32** 如果存在非零解，则图中的曲线必须与表示公式左侧部分的对角线相交——虚线所示。当$\langle k \rangle$较小时，两条线只在$S=0$处相交，表明巨连通分支的相对大小为零。只有当$\langle k \rangle$超过某个阈值时，非零解才会出现。

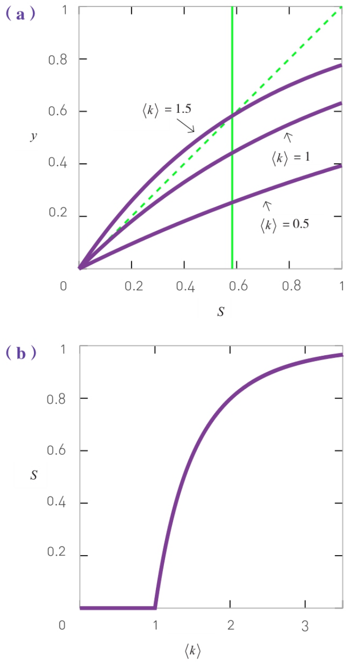


图2-18　图解法

（**a**）三条紫色曲线分别对应$\langle k\rangle$ =0.5、$\langle k\rangle$ =1、$\langle k\rangle$ =1.5时的y =1-exp [-$\langle k\rangle$ k S]。绿色虚线表示对角线y =S 。虚线和紫色曲线的交点对应着公式**2.32** 的解。当$\langle k\rangle$ =0.5时，两条线只有一个交点S =0，表示此时没有巨连通分支。$\langle k\rangle$ =1.5时的曲线在S =0.583处有一个解（绿色垂直线）。$\langle k\rangle$ =1时的曲线对应着临界点——只在S =0处有一个解和存在非零解的分界点。

（**b**）公式**2.32** 所预测的巨连通分支大小关于平均度的函数。[33]

要确定非零解开始出现时的$\langle k\rangle$ 值，我们需要对公式**2.32** 求导。相变点所处的位置是公式**2.32** 左右两侧的导数值相同的地方，即：

$$\frac{d}{dS}\left(1-e^{-\langle k\rangle S}\right)=\langle k\rangle e^{-\langle k\rangle S}=1 \qquad (2.33)$$

令S =0，可以得出相变点发生的位置为$\langle k\rangle$ =1（也可以参考进阶阅读**2.F** ）

2.15

进阶阅读2.D
分支大小

图**2-7** 给出了巨连通分支的大小，并提出了一个重要的问题：给定$\langle k \rangle$ 的情况下，网络中有多少个连通分支？连通分支大小分布如何？本节将讨论这些问题。

分支大小的分布

对于随机网络，一个随机选择的节点属于大小为s 的连通分支（不同于巨连通分支G ）的概率是[33]：

$$p_s \sim \frac{\left(s\langle k \rangle\right)^{s-1}}{s!} e^{-\langle k \rangle s} \qquad (2.34)$$

用$\exp\left[\left(s-1\right)\ln\langle k \rangle\right]$ 代替$\langle k \rangle^{s-1}$ ，使用斯特林公式$s! \approx \sqrt{2\pi s}\left(\frac{s}{e}\right)^s$ ，对于大的s 我们有：

$$p_s \sim s^{-3/2} e^{-\left(\langle k \rangle -1\right)s+\left(s-1\right)\ln\langle k \rangle} \qquad (2.35)$$

因此，分支大小分布由两部分构成：一个缓慢衰减的幂律项$s^{-3/2}$ 和一个快速衰减的指数项$e^{-\left(\langle k \rangle -1\right)s+\left(s-1\right)\ln\langle k \rangle}$ 。当s 较大时，指数项占主导地位，根据公式**2.35** 可以预测，此时大的分支不存在。临界点状态时，$\langle k \rangle = 1$ ，指数项消失了，因此p_s 服从幂律

$$p_s \sim s^{-3/2} \qquad (2.36)$$

由于幂律函数衰减相对较慢，临界点状态时，分支的大小分布可以很宽，和系统处于相变时的性质一致（进阶阅读**2.F** ）。图**2-19** 中的数值模拟结果支持了上述预测。

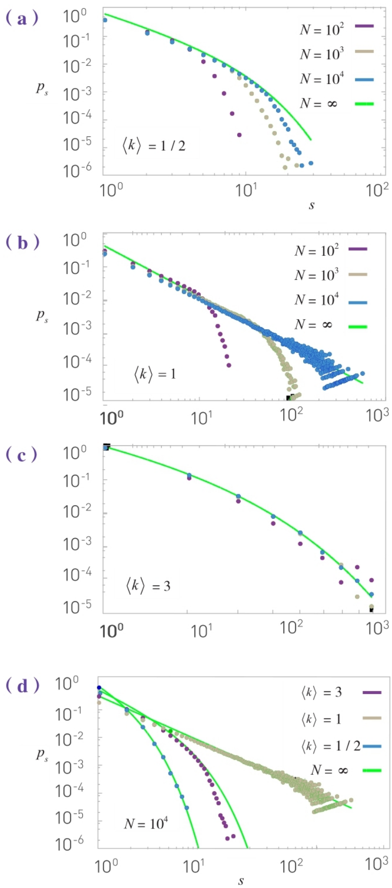


图2-19　分支大小分布

随机网络中的分支大小分布p_s，不包括巨连通分支。

（a）～（c）$\langle k \rangle$和N的不同取值对应的p_s，当N很大时，p_s收敛到公式2.34的预测结果。

（d）$N=10^4$时，不同$\langle k \rangle$对应的p_s。$\langle k \rangle<1$和$\langle k \rangle>1$时，p_s的分布是指数形式。不过，在临界点$\langle k \rangle=1$时，分布服从幂律（公式2.36）。连续的绿线对应公式2.35。关于随机网络中分支大小分布的第一个数值研究是在1998年完成的[34]，该研究激发了人们对复杂网络的研究兴趣。

平均分支大小

计算表明，平均分支大小（不包括巨连通分支）服从：[33]

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle + \langle k \rangle N_G / N} \quad (2.37)$$

当$\langle k \rangle < 1$时，随机网络中不存在巨连通分支（即$N_G = 0$），因此公式**2.37** 变成了

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle} \quad (2.38)$$

当平均度接近临界点$\langle k \rangle = 1$时，公式**2.38** 发散。因此，当我们接近临界点时，分支的大小增加，标志着巨连通分支在$\langle k \rangle = 1$时开始出现。N 较大时的数值模拟结果也支持这些预测（图**2-20**）。

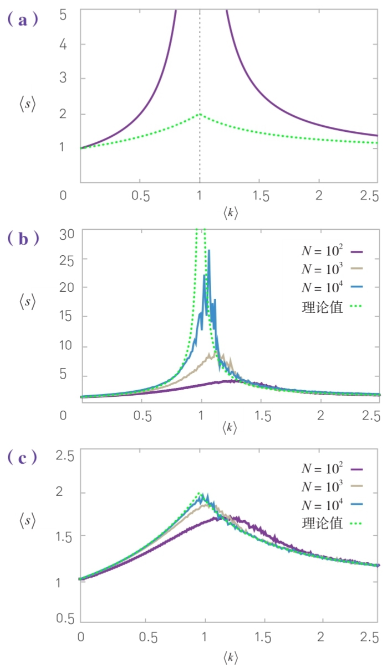


图**2-20**　平均分支大小

（a）公式**2.39** 所预测的，一个随机选择的节点所属分支的平均大小（紫色）。绿线是公式**2.37** 预测的所有分支的平均大小。[33]

（b）随机网络的平均分支大小。随机选择一个节点，得到其所属分支的大小。该度量是有偏差的，因为大小为s 的分支会被数s 次。N 越大，数值模拟结果越符合公式**2.37** 的预测。根据预测，$\langle s \rangle$在临界点$\langle k \rangle = 1$处发散，验证了相变点

的存在（进阶阅读**2.F**）。

（c）随机网络的平均分支大小。通过让每个分支只选择一次，我们纠正了（b）中存在的偏差。N 越大，数值模拟结果越符合公式**2.39** 的预测。

要用公式**2.37** 确定$\langle k\rangle>1$时的平均分支大小，我们需要先计算出巨连通分支的大小。这一点可以通过一种自洽的方式实现，平均分支大小在$\langle k\rangle>1$时随着$\langle k\rangle$的增加而减小，这是因为大多数分支逐渐被巨连通分支吸收。

注意，公式**2.37** 预测的是一个随机选择节点所属分支的大小。这是一个有偏差的度量，这是因为一个节点属于较大分支的概率要高于其属于较小分支的概率。偏差和分支大小s是线性关系。纠正这种偏差之后，我们得到的平均分支大小和我们逐一检查各个分支然后求出其大小的平均值是一样的[33]：

$$\langle s'\rangle=\frac{2}{2-\langle k\rangle+\langle k\rangle N_G/N} \quad (2.39)$$

图**2-20** 为公式**2.39** 提供了数值模拟支持。

2.16 进阶阅读2.E 全连通状态

为了确定大部分节点成为巨连通分支一部分时的$\langle k\rangle$，我们计算出一个随机选择的节点和巨连通分支之间没有链接的概率，这里利用了$N_G\approx N$。因此，这$(1-p)^{N_G}\approx(1-p)^N$种和巨连通分支不相连的节点的期望数目是：

$$I_N=N(1-p)^N=N\left(1-\frac{N\cdot p}{N}\right)^N\approx N\mathrm{e}^{-Np} \quad (2.40)$$

这里用到了如下近似：当n 很大时，$\left(1-\frac{x}{n}\right)^n\approx\mathrm{e}^{-x}$。我们使$p$ 足够

大，大到只有一个节点和巨连通分支不相连。此时，$I_N = 1$。因此，根据公式2.40，p 需要满足$Ne^{-Np} = 1$。因此，在网络即将进入全连通状态时，p 的值是：

$$p = \frac{\ln N}{N} \tag{2.41}$$

从而得到公式2.41。

2.17 进阶阅读2.F 相变

随机网络模型在$\langle k \rangle = 1$处出现巨连通分支，让人联想到物理学和化学中广泛研究过的一个现象——相变[35]。我们来看两个例子：

（1）水冰相变（图2-21a）：在温度比较高的时候，水分子进行扩散运动，水分子之间不断地形成小团队然后又分散开。对水进行冷却，到达零度时，水分子突然就停止了扩散运动，形成有序的晶体结构——冰。

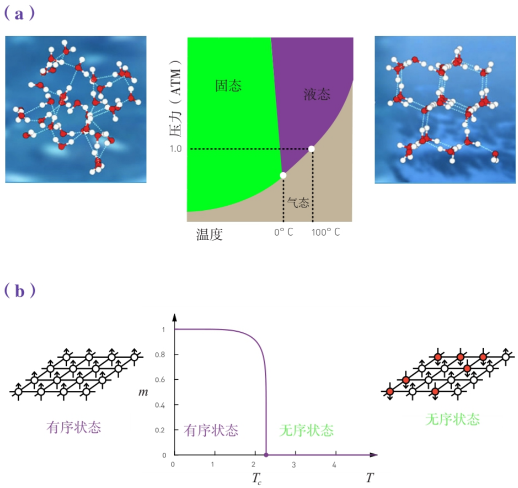


图2-21　相变

（a）水冰相变

将不同水分子连在一起的氢键（虚线）很脆弱，会不断地断裂再形成，维持着部分有序的局部结构（左图）。温度 - 压力相图（中间图）表明，通过降低温度，水会经历一个相变，从液态（紫色）凝结为固态（绿色）。固态水中，每个水分子和4个其他分子紧密连接，形成一个冰晶格（右图）。

（b）磁化相变

在磁性材料中，单个原子的磁矩（自旋）可以指向两个相反的方向。高温下，每个原子随机选择其自旋方向（右图）。处于这种无序状态的系统，其总磁化强度（$m = \Delta M / N$，这里的ΔM是指自旋方向向上的原子数减去自旋方向向下的原子数）为零。相图（中间的图）表明，通过降低温度T，系统在$T = T_c$时经历了一个相变——非零磁化强度开始出现。进一步降低T，m会收敛到1。在这种有序状态下，所有原子的自旋方向相同（左图）。

（2）磁化（图2-21b）：高温下的磁性金属（例如铁）中，每个原子的自旋方向是随机的。到达某个临界温度T_c时，所有原子的自旋方向会变成同一个方向，于是金属变成了磁体。

液体凝固和磁化是相变的两个例子，代表着从无序到有序的转变。相对于冰的完美秩序，液态水是无序的。到达临界温度T_c时，磁体中原本随机的自旋方向被高度有序的共同方向取代。

经历相变的系统有许多性质是普适的。这意味着，这些系统中可以观测到相同的定量模式。例如，岩浆凝固成岩石，陶瓷材料变成超导体。而且，在相变点（也称临界点）附近，许多有趣的量都服从幂律分布。

随机网络在临界点$\langle k \rangle = 1$附近观测到的现象，在很多方面都和相变类似：

（1）图2-7a和图2-21b的磁化之间的相似性并不是偶然的：它们都代表着从无序到有序的转变。在随机网络中，这对应着$\langle k \rangle$超过临界点$\langle k \rangle = 1$时巨连通分支的出现。

（2）接近凝固点时，大小不同的冰晶体形成。到达临界温度时，原子的自旋方向开始分区域变得一致起来，形成

大小不同的磁畴。冰晶体的大小和磁畴的大小都服从幂律分布。与之类似，虽然$\langle k\rangle<1$和$\langle k\rangle>1$时连通分支大小p_s 服从指数分布，在相变点时，p_s 服从幂律分布（公式**2.36** ），表示大小分布很广的连通分支同时存在。

（**3**） 到达临界点时，冰晶体或磁畴的平均大小开始发散，预示着整个系统开始变成一个冰晶体或所有原子的方向开始变得一致。与之类似，在接近$\langle k\rangle=1$时，随机网络中连通分支的平均大小$\langle s\rangle$ 开始发散（图**2-20** ）。

2.18 进阶阅读2.G 小世界的修正

公式**2.18** 是网络直径的一种近似，只在N 非常大且d 较小时有效。实际上，随着$\langle k\rangle^d$ 逐渐接近网络大小N ，$\langle k\rangle^d$ 的增幅一定会降下来。这是因为，网络中没有足够多的节点来支持$\langle k\rangle^d$ 的继续增大。这种有限尺寸效应导致了对公式**2.18** 的修正。对于平均度为$\langle k\rangle$ 的随机网络，对网络直径的更好近似[36] 是：

$$d_{\max}=\frac{\ln N}{\ln\langle k\rangle}+\frac{2\ln N}{\ln\left[-W\left(\langle k\rangle\exp-\langle k\rangle\right)\right]} \quad (2.42)$$

这里，兰伯特函数$W(z)$是$f(z)=z\exp(z)$的逆函数。右侧第一项是公式**2.18** ，第二项是依赖于平均度的修正项。修正项增加了网络直径，解释了这样一个事实：在d 接近网络直径$d_{\max}$ 时，节点数目的增加要比$\langle k\rangle^d$ 慢。在公式**2.42** 的各种极限情况下，修正项的大小会变得更明显。

在$\langle k\rangle\to1$的极限情况下，通过计算兰伯特函数可以求出此时的网络直径[36]

$$d_{\max}=3\frac{\ln N}{\ln\langle k\rangle} \quad (2.43)$$

因此，当巨连通分支出现时，网络直径是公式**2.18** 所预测值的3倍。这是因为，在临界点$\langle k \rangle$ =1处，网络是树状结构，由几乎没有环的长链构成，从而导致d_{max}的增加。

在$\langle k \rangle \to \infty$的极限情况下，此时网络非常稠密，公式**2.42** 变成了：

$$d_{max} = \frac{\ln N}{\ln \langle k \rangle} + \frac{2\ln N}{\langle k \rangle} + \ln N \left(\frac{\ln \langle k \rangle}{\langle k \rangle^2} \right) \quad (2.44)$$

因此，随着$\langle k \rangle$的增加，第二项和第三项消失，公式**2.42** 收敛到公式**2.18** 。

2.19
参考文献

[1] A.-L. Barabási. *Linked: The new science of networks* .Plume Books, 2003.

[2] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)* , 6:290-297, 1959.

[3] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* ., 5:17-61, 1960.

[4] P. Erdös and A. Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist* ., 38:343-347, 1961.

[5] P. Erdös and A. Rényi. On the Strength of Connectedness of a Random Graph, *Acta Math. Acad. Sci. Hungary* , 12: 261-267, 1961.

[6] P. Erdös and A. Rényi. Asymmetric graphs. *Acta Mathematica Acad. Sci. Hungarica* , 14:295-315, 1963.

[7] P. Erdös and A. Rényi. On random matrices. *Publ. Math. Inst. Hung. Acad. Sci* ., 8:455-461, 1966.

[8] P. Erdös and A. Rényi. On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hungary* , 17:359-368, 1966.

[9] P. Erdös and A. Rényi. On random matrices II. *Studia Sci. Math. Hungary* , 13:459-464, 1968.

[10] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics* , 30:1141-1144, 1959.

[11] R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biology* , 13:107-117, 1951.

[12] P. Hoffman. *The Man Who Loved Only Numbers: The Story of Paul* Erdös *and the Search for Mathematical Truth* .Hyperion Books, 1998.

[13] B. Schechter. *My Brain is Open: The Mathematical Journeys of Paul* Erdös. Simon and Schuster, 1998.

[14] G. P. Csicsery. *N is a Number: A Portait of Paul Erdős* , 1993.

[15] B. Bollobás. *Random Graphs* . Cambridge University Press, 2001. [16] L. C. Freeman and C. R. Thompson. Estimating Acquaintanceship.

[16] L. C. Freeman and C. R Volume, pg. 147-158, in *The Small World* , Edited by Manfred Kochen (Ablex, Norwood, NJ), 1989.

[17] H. Rosenthal. *Acquaintances and contacts of Franklin Roosevelt* . Unpublished thesis. Massachusetts Institute of Technology, 1960.

[18] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *ACM Web Science 2012:Conference Proceedings* , pages 45-54. ACM Press, 2012.

[19] R. Albert and A.-L. Barabási. Statistical mechanics of

complex networks. *Reviews of Modern Physics* , 74:47-97, 2002.

[20] I. de Sola Pool and M. Kochen. Contacts and Influence. *Social Networks* , 1: 5-51, 1978.

[21] H. Jeong, R. Albert and A. L. Barabási. Internet: Diameter of the world-wide web. *Nature* , 401:130-131, 1999.

[22] S. Lawrence and C.L. Giles. Accessibility of information on the Web. *Nature* , 400:107, 1999.

[23] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks* , 33:309-320, 2000.

[24] S. Milgram. The Small World Problem. *Psychology Today* , 2: 60-67, 1967.

[25] J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociometry* , 32:425-443, 1969.

[26] K. Frigyes, “Láncszemek,” in *Minden másképpen van* (Budapest: Atheneum Irodai es Nyomdai R.-T. Kiadása, 1929), 85-90. English translation is available in [27].

[27] M. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks* . Princeton University Press, 2006.

[28] J. Guare. *Six degrees of separation* . Dramatist Play Service, 1992.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of‘small-world’ networks. *Nature* , 393: 409-10, 1998.

[30] T. S. Kuhn. T*he Structure of Scientific Revolutions* .University of Chicago Press, 1962.

[31] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[32] A.-L. Barabási, R. Albert, and H. Jeong. Meanfield theory

for scale-free random networks. *Physica A* , 272:173-187, 1999.

[33] M. Newman. *Networks: An Introduction* . Oxford University Press, 2010.

[34] K. Christensen, R. Donangelo, B. Koiller, and K. Sneppen. Evolution of Random Networks. *Physical Review Letters* , 81:2380-2383, 1998.

[35] H. E. Stanley. *Introduction to Phase Transitions and Critical Phenomena* . Oxford University Press, 1987.

[36] D. Fernholz and V. Ramachandran. The diameter of sparse random graphs. *Random Structures and Algorithms* , 31:482-516, 2007.

第3章
无标度性质



图3-0　托马斯·萨拉切诺：艺术与网络

托马斯·萨拉切诺（Tomás　　Saraceno）受蜘蛛网和神经网络的启发创作了这个作品。作为一名建筑师，他拥有从工程学、物理学、化学、航空学及材料学等多个角度洞察世界的能力，将网络作为其思想和灵感的来源。这幅图展示的是在迈阿密艺术博物馆展出的一件作品，是这位艺术家在展现复杂网络方面的一个例子。

本章索引

3.1 本章概览 / **129**

3.2 幂律和无标度网络 / **132**

3.3 枢纽节点 / **137**

3.4 无标度的含义 / **142**

3.5 普适性 / **146**

3.6 超小世界性质 / **151**

3.7 度指数的作用 / **155**

3.8 生成任意度分布的网络 / **159**

3.9 小结 / **165**

3.10 课后习题 / **167**

3.11 进阶阅读**3.A**
幂律分布 / **169**

3.12 进阶阅读**3.B**
绘制幂律分布 / **175**

3.13 进阶阅读**3.C**
估计度指数 / **179**

3.14 参考文献 / **187**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森　**SARAH MORRISON**

加布里埃莱·穆塞拉　**GABRIELE MUSELLA**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

莫罗·马蒂诺　**MAURO MARTINO**

菲利普·赫费尔　**PHILIPP HOEVEL**

罗伯塔·西纳特拉　**ROBERTA SINATRA**

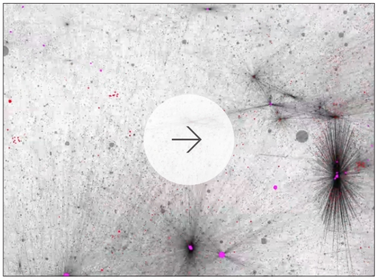
3.1 本章概览

万维网中，节点是网页，节点间的链接被称为统一资源定位符（URL）。借助URL，人们可以通过点击，从一个网页跳转到另一个网页。万维网是目前为止人类建造出的最大网络，其所包含的网页数目估计超过了1万亿（$N \approx 10^{12}$），比人脑中的神经元数目（$N \approx 10^{11}$）还要多。

无论怎么夸大万维网对于人类生活的重要性都不为过。与之类似，无论怎么夸大万维网在网络科学发展过程中的重要作用也都不为过：万维网不仅促进了很多网络基本特性的发现，还是大多数网络指标的标准测试床。

我们可以使用被称为网络爬虫（crawler）的软件来绘制出万维网的连接图。网络爬虫可以从任何一个网页开始，识别出该网页中包含的链接（URL）。然后，网络爬虫下载这些链接所指向的网页，并识别出这些网页中所包含的链接，以此类推。上述过程反复进行，得到万维网的一个局部地图。包括谷歌或必应（Bing）在内的搜索引擎，都是通过网络爬虫发现和索引新网页的，从而维护着一个详细的万维网地图。

首个为理解万维网背后的网络结构而绘制出的万维网地图，出自郑浩雄之手。他绘制的这个万维网地图由nd.edu域名[1]下的30万个网页和150万个链接构成（在线资源**3.1**）。绘制该地图的目的是将万维网的网络性质和随机网络模型进行对比。实际上，在1998年，我们完全有理由相信万维网可以用随机网络很好地近似。每个网页的内容反映了其创建者的个人兴趣或专业兴趣——不论创建者是个人还是组织。由于兴趣的多样性，这些网页上的链接似乎会指向随机选择的网页。



在线资源3.1　透视万维网

该视频展示了对万维网的透视，正是这种方式推动了无标度性质的发现[1]。该网络在表1-1 和图3-1 中都有展示，对于该网络诸多特性的分析贯穿本书。

扫描二维码下载“湛庐阅读”App，回复书名，了解无标度性质是如何被发现的。

对图3-1　的粗略浏览会支持上述观点：在万维网的连接图背后似乎隐藏着大量的随机性。然而，仔细观察就会发现，万维网地图和随机网络之间存在一些令人费解的差异。实际上，在随机网络中，高度连接的节点——枢纽节点，是几乎不可能出现的。相比之下，图3-1 中同时存在大量度很小的节点和少数几个枢纽节点——拥有链接数异常多的节点。

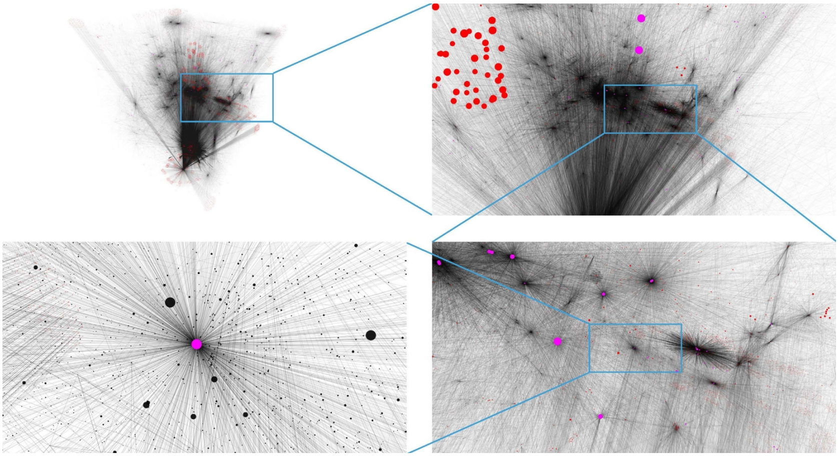


图3-1　万维网的拓扑结构

郑浩雄在1998年绘制完成的万维网地图 [1] 。这一系列图片对万维网中的一个局部区域进行了逐步放大。第一幅图展示了全部325 729个节点，为我们提供了一个全局视角。图中，链接数超过50的节点使用红色标出，链接数超过500的节点用紫色标出。这一系列特写揭示了无标度网络中少量高度连接的节点——枢纽节点的存在。图片来自莫罗·马蒂诺。

在这一章中，我们会看到，枢纽节点并不是万维网所独有的，而是出现在大多数真实网络中。枢纽节点标志着真实网络中蕴含着一种比随机性更深层次的组织原则——我们称之为无标度性质。因此，我们对真实网络的度分布进行探索，从而得以揭示和刻画无标度网络。本章所讨论的解析和实证结果，是本书剩余部分关于网络建模的基础。实际上，我们将看到，无论研究网络的哪种性质，例如社区结构和传播过程，都必须在掌握了网络度分布的情况下进行。

3.2 幂律和无标度网络

假如万维网是一个随机网络，那么万维网中节点的度应该服从泊松分布。然而，如图**3-2**　所示，泊松分布并不适用于拟合万维网的度分布。相反，在双对数坐标下，万维网度分布近似形成了一条直线。这表明，万维网的度分布可以近似为：

$$p_k \sim k^{-\gamma} \tag{3.1}$$

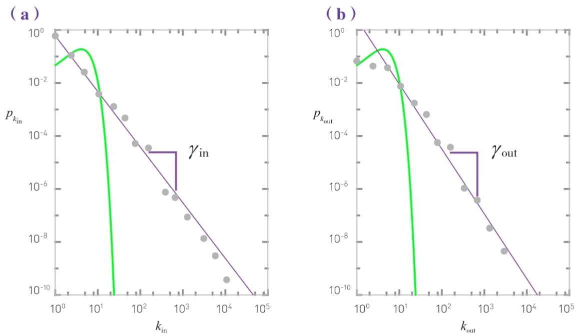


图3-2　万维网的度分布

万维网的（a）入度分布和（b）出度分布，此分布由阿尔伯特等人在1999年的研究中绘制完成 [1] 。注意，度分布是在双对数坐标（log-log图）下展示的，双对数坐标下幂律分布是一条直线。图中的点代表实测数据，直线代表对实测数据的幂律分布拟合，度指数分别为$\gamma_{in}=2.1$和$\gamma_{out}=2.45$。另外，绿色线给出了使用泊松函数对度分布进行拟合的结果，拟合时使用了从万维网中实测得到的平均度$\langle k_{in}\rangle=\langle k_{out}\rangle=4.60$。

公式3.1 被称为幂律分布。本书中，指数γ 被称为度指数（边栏3.1 ）。对公式3.1 取对数，我们可以得到：

边栏3.1

80/20法则和前1%

经济学家维弗雷多·帕累托（Vilfredo Pareto，见图3-3 ）发现，在意大利，极少数富人占有着大部分财产，而人口的大多数却只占有小部分财产。他将这种差异和观测到的一个现象联系在一起：人口收入服从幂律分布。这是关于幂律分布的第一个有记载的报道 [3] 。他的这一发现广泛流传，被称为80/20法则：20%的人口拥有80%的财产。



图3-3 维弗雷多·帕累托（1848—1923）

意大利经济学家、政治学家和哲学家。帕累托在理解收入分布和个人选择分析方面做出了重要贡献。很多基本原理都是以他的名字命名的，例如帕累托效率（Pareto efficiency）、帕累托分布（Pareto distribution，幂律分布的别名）和帕累托法则（也称80/20法则）。

80/20法则在很多领域都有所体现。例如，在管理学领域中，80%的利润由20%的员工创造。与之类似，在会议中，80%的决策是在20%的时间内做出的。

80/20法则在网络中也同样存在。万维网中，80%的链接指向了15%的网页；科学引文网络中，80%的引用指向38%的科学家；好莱坞演员网络中，80%的链接来自30%的演员 [4] 。实际上，大部分符合幂律分布的变量都遵循80/20法则。

在2009年的经济危机中，幂律分布又有了新的意义。“占领华尔街运动”揭露了这样一个事实：在美国，1%的人口拥有美国全部财产的15%。这种“1%现象”是收入差异的一个缩影，是收入服从幂律分布的结果。

$$\log p_k \sim -\gamma \log k \quad (3.2)$$

如果公式**3.1** 成立，则$\log p_k$ 和$\log k$ 呈线性关系，斜率即为度指数γ（图**3-2** ）。

由于万维网是一个有向网络，每个网页可以用出度k_{out} 和入度k_{in} 来表示。其中，出度k_{out} 表示从该网页指向其他网页的链接数，入度k_{in} 表示从其他网页指向该网页的链接数。因此，我们必须区分这两种不同的度分布：$p_{k_{out}}$ 表示一个随机选择的网页其出度为k_{out} 的概率，$p_{k_{in}}$ 表示一个随机选择的网页其入度为k_{in} 的概率。在万维网这个例子中，$p_{k_{in}}$ 和$p_{k_{out}}$ 都可以近似为幂律分布：

$$p_{k_{in}} \sim k^{-\gamma_{in}} \quad (3.3)$$

$$p_{k_{out}} \sim k^{-\gamma_{out}} \quad (3.4)$$

其中，γ_{in} 和γ_{out} 分别是入度分布和出度分布的度指数（图**3-2**）。一般而言，γ_{in} 和γ_{out} 可以不相同。例如，在图**3-1** 中，我们有$\gamma_{in} \approx 2.1$和$\gamma_{out} \approx 2.45$。

图**3-2** 所示的实证结果表明存在这样一种网络，其度分布和随机网络的泊松分布非常不同。我们将这类网络称为无标度网络，定义如下 [2] ：无标度网络是度分布服从幂律分布的网络。

如图**3-2** 所示，对万维网而言，幂律分布的跨度几乎多达4个数量级，这也是我们将其称为无标度网络的原因。在这种情况下，无标度性质同时适用于入度和出度。

为了更好地理解无标度性质，我们需要用更精确的术语来定义幂律分布。因此，接下来我们将讨论本书中经常使用的离散和连续形式的幂律分布。

离散形式

由于节点的度是正整数，$k = 0,1,2,\dots$，因此，表示一个节点正好有k 个链接的概率p_k 是一种离散形式的幂律分布：

$$p_k = C k^{-\gamma} \quad (3.5)$$

常数C 由如下归一化条件来确定：

$$\sum_{k=1}^{\infty} p_k = 1 \quad (3.6)$$

由公式**3.5** 可以得到：

$$C \sum_{k=1}^{\infty} k^{-\gamma} = 1$$

因此，

$$C=\frac{1}{\sum_{k=1}^{\infty}k^{-\gamma}}=\frac{1}{\zeta(\gamma)} \tag{3.7}$$

其中，$\zeta(\gamma)$是黎曼泽塔函数。因此，对于任意$k>0$，离散形式的幂律分布为：

$$p_k=\frac{k^{-\gamma}}{\zeta(\gamma)} \tag{3.8}$$

注意，公式**3.8** 在$k=0$时发散。需要的时候，我们可以单独定义p_0，用来表示没有链接的节点所占的比例。在这种情况下，计算公式**3.7** 中的C时需要把p_0考虑进来。

连续形式

方便起见，在进行解析计算时，通常会假设度可以是任意正实数。在这种情况下，我们可以把幂律度分布写成：

$$p(k)=Ck^{-\gamma} \tag{3.9}$$

使用归一化条件

$$\int_{k_{\min}}^{\infty}p(k)dk=1 \tag{3.10}$$

可以得到

$$C=\frac{1}{\int_{k_{\min}}^{\infty}k^{-\gamma}dk}=(\gamma-1)k_{\min}^{\gamma-1} \tag{3.11}$$

因此，度分布的连续形式为：

$$p(k)=(\gamma-1)k_{\min}^{\gamma-1}k^{-\gamma} \tag{3.12}$$

其中，$k_{\min}$是幂律分布（公式**3.8**）保持成立的最小度。

注意，离散形式中的p_k有明确的含义：一个随机选择的节点其度为k的概率。相比之下，连续形式中具有物理意义的只有$p(k)$的积

分。例如，一个随机选择的节点其度介于k_1和k_2之间的概率为：

$$\int_{k_1}^{k_2} p(k)\,dk \tag{3.13}$$

综上所述，度分布服从幂律分布的网络被称为无标度网络。如果网络是有向的，无标度性质分别适用于入度和出度。为了便于从数学角度研究无标度网络的性质，我们介绍了离散形式和连续形式的幂律分布。需要注意的是，无标度性质与采用离散形式还是连续形式无关。

3.3 枢纽节点

随机网络和无标度网络的主要区别体现在度分布的尾部，即p_k 中k 比较大的区域。为了说明这一点，我们在图**3-4** 中将幂律分布和泊松分布进行了对比。我们发现：

（**1**） 当k 比较小时，幂律分布在泊松分布之上。这表明，无标度网络中有大量度比较小的节点，而随机网络中这样的节点很少出现。

（**2**） 当k 处于$\langle k\rangle$附近时，泊松分布在幂律分布之上。这表明，随机网络中大多数节点的度在$\langle k\rangle$附近。

（**3**） 当k 比较大时，幂律分布再次位于泊松分布之上。如果将p_k 画在双对数坐标下（图**3-4b** ），这种差异会格外明显。这一现象表明，在无标度网络中观察到一个大度节点——枢纽节点的概率，要比在随机网络中观察到这样一个节点的概率大好几个数量级。

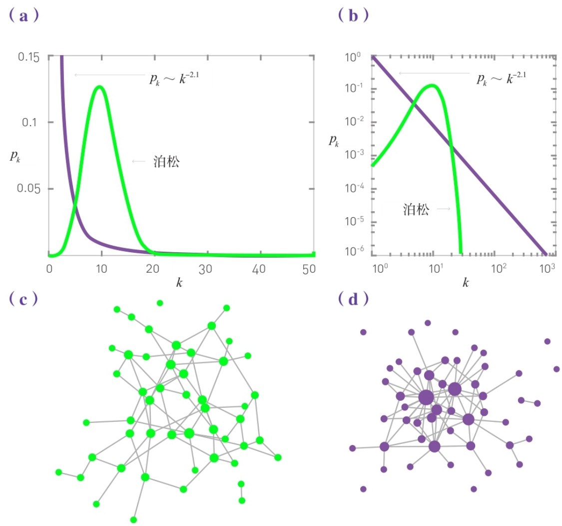


图3-4　泊松分布和幂律分布

（a）在线性坐标下的泊松函数和幂律函数（$\gamma=2.1$）。两个度分布的平均度均为$\langle k\rangle=11$。

（b）与（a）相同，但坐标换成了双对数坐标，以便观察两个函数在k比较大的区域上的差异。

（c）一个$\langle k\rangle=3$、$N=50$的随机网络。从图中可以看出，随机网络中大多数节点的度都在平均度附近，即$k\approx\langle k\rangle$。

（d）一个$\gamma=2.1$、$\langle k\rangle=3$的无标度网络。从图中可以看出，无标度网络中大量小度节点和少量拥有很多链接的枢纽节点共存。图中每个节点的大小和它的度成正比。

让我们以万维网为例来阐述上述差异的量级。假如其度分布服从泊松分布，则观察到一个$k=100$的节点的概率是$p_{100}\approx10^{-94}$；而当p_k服从幂律分布时，这个概率为$p_{100}\approx4\times10^{-4}$。因此，假如万维网是一个平均度为$\langle k\rangle=4.6$、节点数为$N\approx10^{12}$的随机网络，则有：

$$N_{k\geq100}=10^{12}\sum_{k=100}^{\infty}\frac{(4.6)^k}{k!}\mathrm{e}^{-4.6}\simeq10^{-82}\qquad(3.14)$$

这意味着，链接数不少于100的节点几乎没有。相反，在已知万维网的度分布服从幂指数$\gamma_{in}=2.1$的幂律分布的情况下，我们有N_k

$N_{\geq 100}=4\times 10^9$。也就是说，度$k\geq 100$的节点数不少于40亿。

最大的枢纽节点

所有真实网络的大小都是有限的。万维网估计有$N\approx 10^{12}$个节点；全球社交网络的大小为$N\approx 7\times 10^9$。这些数字虽然很大，但仍然是有限的。相比之下，其他网络的大小就更加黯然失色了：人类细胞中的基因网络大约拥有2万个基因；大肠杆菌的代谢网络中只有大约1 000个节点。我们不禁会问：网络的大小会如何影响其枢纽节点的大小呢？为了回答这个问题，我们需要计算出最大度k_{max}，k_{max}也被称为度分布p_k的自然截断（natural cutoff）。最大度k_{max}表示网络中最大枢纽节点的期望大小。

为了得到一些启发，我们先对指数分布进行上述计算：

$$p(k)=Ce^{-\lambda k}$$

对于最小度为k_{min}的网络，由归一化条件

$$\int_{k_{min}}^{\infty} p(k)\,dk=1 \tag{3.15}$$

可以得出$C=\lambda e^{\lambda k_{min}}$。为了计算$k_{max}$，我们假设大小为$N$的网络中最多有一个节点在范围$(k_{max},\infty)$内（进阶阅读**2.B**）。换句话说，假设观察到有一个节点的度超过k_{max}的概率为$1/N$：

$$\int_{k_{min}}^{\infty} p(k)\,dk=\frac{1}{N} \tag{3.16}$$

由公式**3.16**可以得到：

$$k_{max}=k_{min}+\frac{\ln N}{\lambda} \tag{3.17}$$

由于$\ln N$关于系统大小N增长得很缓慢，因此公式**3.17**告诉我们，最大度k_{max}和最小度k_{min}之间的偏差并不明显。当度分布为泊松分布时，上述计算较为复杂，但得到的k_{max}关于N的依赖关系甚至比公式**3.17**中的对数依赖增长得更慢（进阶阅读**2.B**）。

对于无标度网络，根据公式**3.12** 和公式**3.16** ，可以得出自然截断为：

$$k_{max}=k_{min}N^{\frac{1}{\gamma-1}} \tag{3.18}$$

因此，网络越大，其最大枢纽节点的度就越大。k_{max} 对于N 的多项式依赖意味着，在大型无标度网络中，最大度k_{max} 和最小度k_{min} 之间可能存在几个数量级的差异（图**3-5** ）。

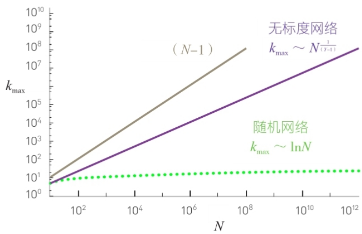


图**3-5** 无标度网络中枢纽节点的度很大

无标度网络和随机网络中对节点最大度（自然截断）的估计。两个网络的平均度均为$\langle k\rangle=3$。对于无标度网络，我们选择$\gamma=2.5$。为了对比，我们还在图中画出了线性变化曲线$k_{max}\sim N-1$，即网络全连接的情况。总体而言，对于同样的N 和$\langle k\rangle$，无标度网络中枢纽节点的度比随机网络中最大节点度要大几个数量级。

为了说明指数度分布的网络和无标度网络中最大度的不同，让我们回到图**3-1** 所示的包含$N\approx3\times10^5$ 个节点的万维网。该网络中，$k_{min}=1$。如果度分布服从指数函数分布，根据公式**3.17** 可以得出，在$\lambda=1$时网络的最大度应为$k_{max}\approx14$。对于大小相似、$\gamma=2.1$的无标度网络，根据公式**3.18** 可以得出，网络的最大度为$k_{max}\approx95000$，和指数分布相比有明显不同。注意，图**3-1** 中的万维网中最大入度为10 721，与无标度网络预测的k_{max} 量级相当。这再次印证了我们的结论：枢纽节点在随机网络中不存在，在无标度网络却自然出现。

综上所述，随机网络和无标度网络的主要区别，从根本上来说是由于泊松分布和幂律分布的形状不同而造成的。随机网络中，大多数节点具有相近的度，因此枢纽节点是不存在的。在无标度网络中，枢纽节点不仅可以出现，而且必然出现（图**3-6** ）。此外，无标度网络

拥有的节点数越多，其枢纽节点越大。事实上，在无标度网络中，枢纽节点的大小随着网络大小呈多项式增长，因此可以增长到非常大。相比之下，随机网络中节点最大度的大小随着N呈对数增长或者增长更慢。这意味着即使在非常大的随机网络中，节点最大度依然会很小。

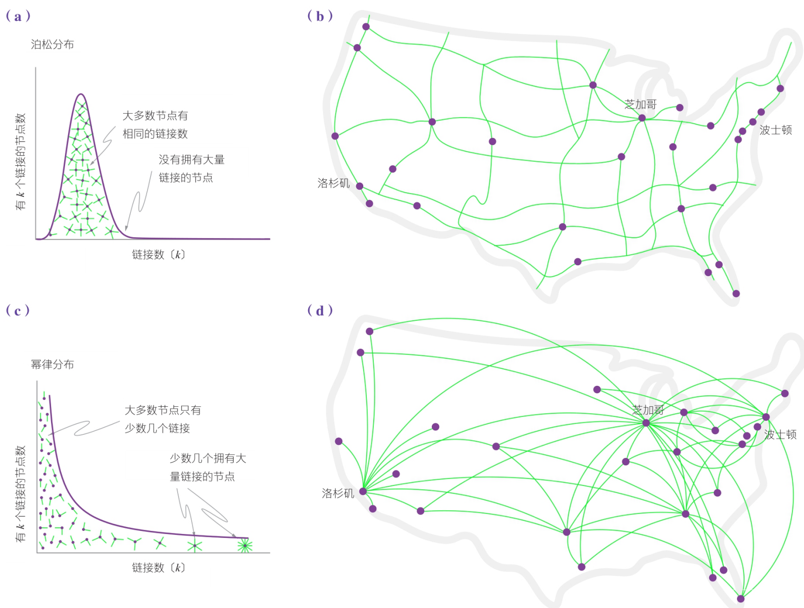


图3-6　随机网络和无标度网络

（**a**）　随机网络的度服从泊松分布，非常像钟形曲线。因此，大多数节点的度都很相近，且不存在具有大量链接的节点。

（**b**）　随机网络看起来有点像全国高速公路网，节点是城市，链接表示主要的高速公路。不会出现拥有上百条高速公路的城市，也几乎没有与高速公路系统不相连的城市。

（**c**）　在具有幂律度分布的网络中，大多数节点都只有少数几个链接。这些大量的小节点通过少数几个拥有大量链接的节点连接在一起。

（**d**）　无标度网络看起来像航空交通网络，节点是飞机场，链接是飞机场之间的直飞航班。大多数飞机场都很小，只有几个航班。然而，我们也有少量非常大的机场，像芝加哥或洛杉矶的机场，它们连接着许多小飞机场，扮演着枢纽节点的角色。

一旦枢纽节点出现，它们将改变我们遍历网络的方式。例如，如果我们要开车从

波士顿去洛杉矶，就必须经过许多城市。但在航空网络中，我们只通过某个枢纽节点，如芝加哥，就能到达大多数目的地。[4]

3.4
无标度的含义

“无标度”一词源于统计物理学的分支——相变理论。相变理论在20世纪六七十年代对幂律进行了广泛且深入的研究（进阶阅读**2.F**）。为了充分理解“无标度”一词的含义，我们需要先熟悉一下度分布的矩。

度分布的n 阶矩定义为：

$$\langle k^n \rangle = \sum_{k_{\min}}^{\infty} k^n p_k \approx \int_{k_{\min}}^{\infty} k^n p(k) dk \qquad (3.19)$$

度分布的低阶矩具有一些重要解释：

（**1**）$n=1$：一阶矩是平均度$\langle k \rangle$。

（**2**）$n=2$：二阶矩$\langle k^2 \rangle$ 可以帮助我们计算方差$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$。方差刻画了度分布的散布程度，其平方根$\sigma$ 被称为标准差。

（**3**）$n=3$：三阶矩$\langle k^3 \rangle$ 体现了度分布的偏斜度，反映了度分布p_k 在平均度$\langle k \rangle$ 附近的对称程度。

对于无标度网络而言，度分布的n 阶矩为：

$$\langle k^n \rangle = \int_{k_{\min}}^{k_{\max}} k^n p(k) dk = C \frac{k_{\max}^{n-\gamma+1} - k_{\min}^{n-\gamma+1}}{n-\gamma+1} \qquad (3.20)$$

通常情况下$k_{\min}$ 是固定的，不过最大枢纽节点的度$k_{\max}$ 会依据公式**3.18** 随系统大小的增大而增大。因此，为了理解$\langle k^n \rangle$，我们需要研究渐进极限$k_{\max} \to \infty$ 情况下的公式**3.20**，从而理解大网络的性质。在该极限情况下，公式**3.20** 预测，$\langle k^n \rangle$ 的值取决于n 和γ 的相互影响：

（1） 如果$n-\gamma+1\leq 0$，则公式3.20 等号右边的第一项$k_{max}^{n-\gamma+1}$ 会随着k_{max}的增加趋近于0。因此，$n\leq\gamma-1$时度分布的矩是有限的。

（2） 如果$n-\gamma+1>0$，则$\langle k^n\rangle$ 在$k_{max}\to\infty$时趋向无穷大。因此，$n>\gamma-1$时度分布的矩是发散的。

对于很多无标度网络，度指数γ 都在2和3之间（表3-1 ）。因此，对于这些网络，在$N\to\infty$的极限下，一阶矩$\langle k\rangle$ 是有限的，但二阶矩$\langle k^2\rangle$ 及更高阶的矩都趋向无穷大。高阶矩的发散可以帮助我们理解“无标度”一词的由来。实际上，如果度服从正态分布，则一个随机选择的节点的度所处的范围通常为

$$k=\langle k\rangle\pm\sigma_k \qquad (3.21)$$

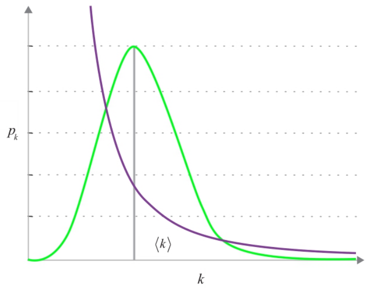
然而，平均度$\langle k\rangle$ 和标准差σ_k 在随机网络和无标度网络中的量级差异很大：

- 随机网络有标度

对于度服从泊松分布的随机网络而言，标准差$\sigma_k=\langle k\rangle^{1/2}$ 总是小于平均度$\langle k\rangle$ 。网络中节点的度在范围$k=\langle k\rangle\pm\langle k\rangle^{1/2}$ 内。换句话说，随机网络中，节点的度相差不大，平均度$\langle k\rangle$ 可以作为随机网络的“标度”。

- 无标度网络没有标度

对于度服从$\gamma<3$的幂律分布的网络而言，一阶矩是有限的，但二阶矩无限大。N 较大时，$\langle k^2\rangle$ （以及σ_k ）的发散表明，度和平均度的偏差可以是任意大。这意味着随机选择一个节点时，我们不知道该节点的度会有多大：该节点的度可以很小，也可以是任意大。因此，当幂律度分布满足$\gamma<3$时，网络中没有有意义的内在标度，这样的网络被称为是“无标度”的（图3-7 ）。



随机网络

随机选择的节点：$k = \langle k\rangle \pm \langle k\rangle^{1/2}$

标度：$\langle k\rangle$

无标度网络

随机选择的节点：$k = \langle k\rangle \pm \infty$

标度：无

图3-7　缺少内在标度

对于任意具有指数限界的分布，如泊松分布或高斯分布，一个随机选择的节点的度会在$\langle k\rangle$附近。因此，$\langle k\rangle$可以作为网络的标度。而对于幂律分布而言，其二阶矩是发散的，一个随机选择的节点的度可以和$\langle k\rangle$相差非常大。因此，$\langle k\rangle$并不能作为内在标度。由于具有幂律度分布的网络缺少内在标度，我们称其为无标度网络。

例如，万维网的平均度是$\langle k\rangle = 4.60$（表3-1）。该网络中$\gamma \approx 2.1$，因此二阶矩是发散的。这意味着在$N \to \infty$的极限下，万维网中随机选择的一个节点，其入度预期为$k = 4.60 \pm \infty$。也就是说，一个随机选择的网页，很可能是度为1或2的节点，因为74.02%的节点的入度都比$\langle k\rangle$小；不过，该网页也可以是拥有数亿链接的节点，例如google.com或facebook.com。

表3-1　真实网络中的度波动

网络	N	L	$\langle k\rangle$	$\langle k_{in}^2\rangle$	$\langle k_{out}^2\rangle$	$\langle k^2\rangle$	γ_{in}	γ_{out}	γ
互联网	192 244	609 066	6.34	—	—	240.1	—	—	3.42*
万维网	325 729	1 497 134	4.60	1546.0	482.4	—	2.00	2.31	—
电网	4 941	6 594	2.67	—	—	10.3	—	—	Exp.
手机通话网络	36 595	91 826	2.51	12.0	11.7	—	4.69*	5.01*	—
电子邮件网络	57 194	103 731	1.81	94.7	1163.9	—	3.43*	2.03*	—
科学合作网络	23 133	93 439	8.08	—	—	178.2	—	—	3.35*
演员网络	702 388	29 397 908	83.71	—	—	47 353.7	—	—	2.12*
引文网络	449 673	4 689 479	10.43	971.5	198.8	—	3.03**	4.00*	—
代谢网络	1 039	5 802	5.58	535.7	396.7	—	2.43*	2.90*	—
蛋白质相互作用网络	2 018	2 930	2.90	—	—	32.3	—	—	2.89*

表中展示了10个参考网络的一阶矩$\langle k\rangle$和二阶矩$\langle k^2\rangle$（有向网络中是$\langle k_{in}^2\rangle$和$\langle k_{out}^2\rangle$）。注意，对于有向网络有$\langle k\rangle=\langle k_{in}\rangle=\langle k_{out}\rangle$。此外，表中还列出了使用进阶阅读**3.A**中所讨论的方法估计出的度指数γ。数字旁的星号表示了度分布拟合的置信度。也就是说，*表示幂律（$k^{-\gamma}$）拟合是统计上可信的，**表示具有指数截断的幂律拟合（公式**3.39**）是统计上可信的。注意，电网不是无标度的；对于该网络，指数形式（$e^{-\lambda k}$）的度分布拟合是统计上可信的，因此我们在该网络对应的最后一栏写了“Exp”。

严格来讲，只有在$N\to\infty$的极限下$\langle k^2\rangle$才会发散。发散对于大小有限的网络也是有意义的。为了说明这一点，表**3-1**列举了10个真实网络的$\langle k^2\rangle$；同时，图**3-8**展示了这些网络的标准差σ。在这些网络中，大多数网络的σ都比$\langle k\rangle$大，意味着不同节点的度相差较大。例如，在表**3-1**列出的万维网中，一个随机选择的节点的入度是$k_{in}=4.60\pm1\ 546$，再次表明了平均度无意义。

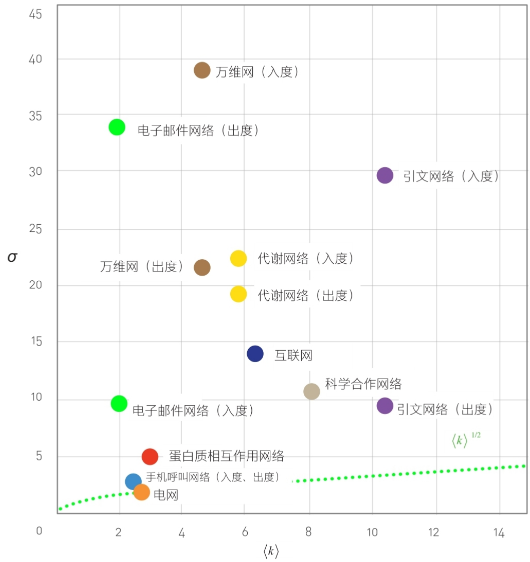


图3-8　真实网络中度分布的标准差很大

对于随机网络，度分布的标准差为$\sigma = \langle k \rangle^{1/2}$，如图中绿色虚线所示。图中的点对应着9个参考网络的σ值，由表**3-1**中的参数计算得到。由于演员网络的$\langle k \rangle$和σ非常大，本图中省略了该网络。对于每个网络，σ的值都比具有相同$\langle k \rangle$的随机网络中预期的σ值要大。唯一的例外是电网，它不是无标度网络。手机呼叫网络虽然是无标度的，但它的γ值非常大，导致它可以用随机网络很好地近似。

总之，“无标度”一词体现的是“缺少内在标度”，这是网络中度相差很大的节点同时存在的结果。这一特性将无标度网络和规则网络或随机网络区分开来。在规则网络中，所有节点具有完全相同的度（$\sigma=0$）；在随机网络中，所有节点的度只在很小的范围内变化（$\sigma=\langle k \rangle^{1/2}$）。在接下来的章节中我们将会看到，无标度网络度分布蕴含的发散性是无标度网络中许多有趣性质的根源——包括网络在节点随机失效时呈现出的健壮性、病毒在无标度网络上的快速传播。

3.5
普适性

尽管万维网和互联网这两个词在媒体中经常被混用，但二者所指

的是完全不同的系统。万维网是一个信息网络，它的节点是网页，链接是URL。而互联网是一个基础设施网络，它的节点是被称为路由器的计算机，链接是指计算机之间的物理链路，例如电缆、光缆或无线链路。

二者之间的区别有着重要的影响：将一个位于波士顿的网页链接到同一台计算机上的其他网页和将其链接到位于布达佩斯的一台计算机上的网页，代价一样大。相比之下，如果我们要在波士顿和布达佩斯的路由器之间建立一个互联网链接，则需要在北美和欧洲之间铺设一条电缆，而这是非常昂贵的。尽管万维网和互联网之间存在这些差异，但这两个网络的度分布都可以用幂律分布很好地近似[1]，[5]，[6]。我们可以从图**3-9**　中看出互联网的无标度特性：少量度较大的路由器将大量只有几个链接的路由器连接在一起。

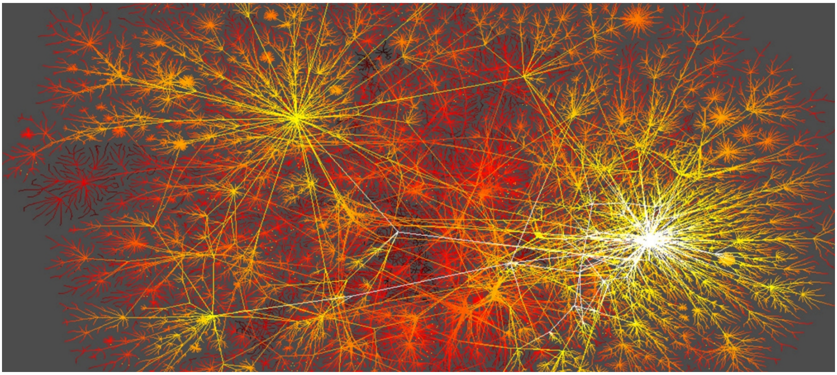


图**3-9**　互联网的拓扑结构

图中所示的是21世纪初的互联网的拓扑结构，该图由CAIDA绘制。CAIDA是加州大学圣迭戈分校的一个机构，致力于收集、分析和描绘互联网数据。这幅互联网地图展示了互联网的无标度性质：少量高度连接的枢纽节点将许多小节点连接在一起。

在过去的10年中，许多具有科学、技术和社会价值的真实网络都被发现具有无标度性质。图**3-10**　展示了许多真实网络的度分布：基础设施网络（互联网），生物网络（蛋白质相互作用网络）、通信网络（电子邮件网络），以及描述学术交流的网络（引文网络）。每个网络的度分布都明显偏离泊松分布，却都能被幂律分布很好地近似。

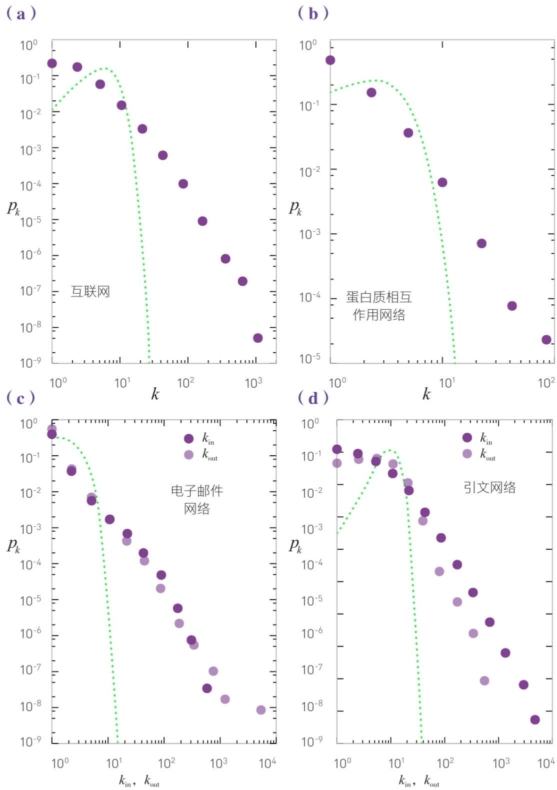


图3-10　许多真实网络是无标度的

表3-1 中列出的4个网络的度分布：

（a）路由器级的互联网；

（b）蛋白质相互作用网络；

（c）电子邮件网络；

（d）引文网络。

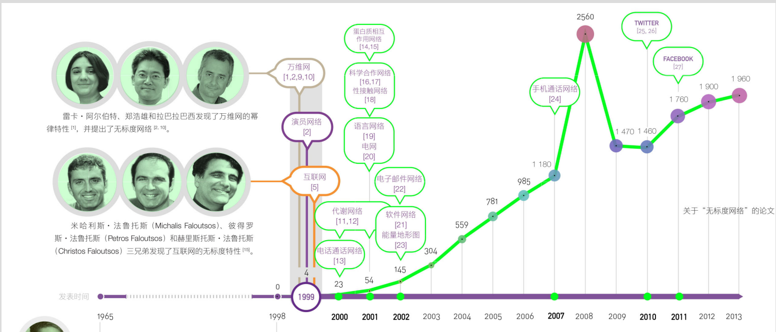
在每个子图中，绿色虚线表示和真实网络具有相同$\langle k \rangle$的泊松分布。从中可以看出，随机网络模型并不能很好地解释观察到的p_k。对有向网络而言，我们将其入度分布和出度分布分开进行展示。

具有无标度性质的系统十分多元化（边栏3.2）。实际上，万维网是一个只有不到20年历史的人造网络，而蛋白质相互作用网络是40亿年进化的产物。一些网络的节点是分子，而另一些网络的节点是计算机。正是这种多元化促使我们将无标度性质视为一种普适的网络特

征。

边栏3.2

无标度网络的时间表



我们预期，在我们拥有详细数据的所有系统中观测到的这种标度不变性不只适用于我们所研究的这些例子，还是很多复杂网络的普遍性质。

——巴拉巴西和阿尔伯特，1999

从研究人员的角度来看，一个关键的问题是：我们如何知道一个网络是不是无标度网络呢？一方面，看一眼度分布就可以立即说出一个网络是不是无标度的：无标度网络中，最小节点的度和最大节点的度明显不同，通常相差几个数量级。相反，在随机网络中，这些节点的度相差不大。由于度指数在预测各种网络性质中起到了非常重要的作用，因此我们需要可以拟合度分布p_k 以及估计出度指数γ 的工具。这促使我们去解决一些绘制和拟合幂律分布方面的问题：

绘制度分布

本章所示的度分布均绘制在双对数坐标下。这样做的主要原因是，当节点的度相差很大时，线性坐标下难以将度分布全部展现出来。为了得到本书中所展现的那种清晰的度分布图，我们使用了对数分级的方式，确保每个数据点对应足够多的观测。绘制网络度分布的实用技巧，可参见进阶阅读**3.B** 。

测量度指数

在双对数坐标下，我们可以通过用直线去拟合p_k，从而快速估计出幂律度分布的度指数。然而，这种方法存在系统偏差，因此会得到不正确的γ。关于估计γ的工具，可参见进阶阅读**3.C**。

真实网络中p_k的形状

许多真实网络中观察到的度分布，和纯幂律分布有一定的偏差。这些偏差可以归因于数据的不完整或者数据收集时的偏差。同时，这些偏差也包含了关于这些网络形成过程的重要信息。我们将在进阶阅读**3.B**中讨论其中的一些偏差，并在第**5**章中探讨这些偏差的起源。

综上所述，自从1999年发现了万维网的无标度性质后，大量具有科学、技术意义的真实网络都被发现具有无标度性质。这些真实网络多种多样，包括生物网络、社会网络及语言网络（边栏**3.2**）。但这并不意味着所有的网络都是无标度的。实际上，许多重要的网络——包括电网和材料科学中观察到的网络，都不具备无标度性质（边栏**3.3**）。

边栏**3.3**

并非所有网络都是无标度的

无标度性质的普遍性并不意味着所有真实网络都是无标度的。相反，一些重要的网络并不具有这一性质：

（**1**） 材料科学中的一些网络——描述晶体或非晶体材料中原子间连接关系的网络。这些网络中，所有节点具有完全相同的度，由化学性质决定（图**3-11**）；

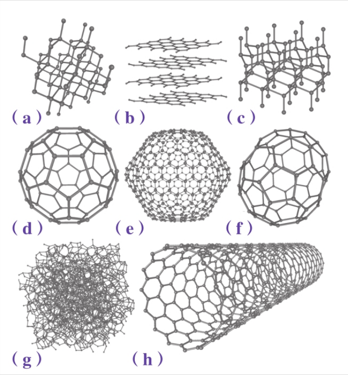


图3-11　材料网络

一个碳原子只能和其他原子共享4个电子，所以无论我们如何放置原子间的相对位置，所得到的网络中每个节点都不可能拥有超过4个链接。因此，在这些网络中，枢纽节点是不存在的，无标度性质也不会出现。上图展示了几种碳的同素异形体（由碳元素组成但网络组织结构不同的材料）。这些不同的结构形成了物理性质和电子性质不同的材料，例如（a）金刚石；（b）石墨；（c）蓝丝黛尔石（lonsdaleite）；（d）C60（足球烯）；（e）C540（一种富勒烯）；（f）C70（另一种富勒烯）；（g）非晶碳；（h）单壁碳纳米管。

（**2**）秀丽隐杆线虫的神经元网络[28]；

（**3**）节点为发电机和交换器、链接为传输电线的电网。

无标度性质的存在，要求节点具有连接任意数量的其他节点的能力。这些链接不需要同时发生：我们不会同时和每个熟人交谈，细胞中的蛋白质也不会同时和其所有的潜在交互对象相互作用。在一些系统中，一个节点能够拥有的链接数是受限的，从而约束了枢纽节点的大小，导致无标度性质不再存在。这种限制在材料科学中很常见（图**3-11**），这解释了为什么材料科学的网络不具备无标度拓扑结构。

3.6 超小世界性质

无标度网络中枢纽节点的出现，引发了一个有趣的问题：枢纽节

点会影响小世界性质吗？图3-4 表明，枢纽节点的确会影响小世界性质：航空公司正是通过建造枢纽节点来减少机场间转机次数的。计算结果也支持上述预期：同等条件下，无标度网络中节点间的平均距离比随机网络中节点间的平均距离要小。

平均距离$\langle d\rangle$ 对系统大小N 和度指数γ 的依赖可以表述为下面的公式[29]，[30]：

$$\langle d\rangle \sim \begin{cases} \text{const.} & \gamma = 2 \\ \ln\ln N & 2 < \gamma < 3 \\ \dfrac{\ln N}{\ln\ln N} & \gamma = 3 \\ \ln N & \gamma > 3 \end{cases} \quad (3.22)$$

接下来，我们将讨论$\langle d\rangle$ 在公式3.22 预测的4种情况下的表现（图3-12）：

异常状态（γ =2）

根据公式3.18 ，当γ =2时，最大枢纽节点的度随系统大小成线性增长，即$k_{\max} \sim N$ 。此时，网络呈现出以一个枢纽节点为中心的星型结构。所有节点相距都很近，因为这些节点都与同一个中心枢纽节点相连。在这种状态下，平均路径长度与N 无关。

超小世界（2<γ <3）

公式3.22 预测，该状态下网络平均距离会按照$\ln\ln N$ 增长，比随机网络中推导出的$\ln N$ 要慢得多。我们将处于该状态下的网络称为超小（ultra-small）网络——枢纽节点通过连接大量的小度节点使彼此之间距离变短，从而大大缩短了路径长度[29] 。为了阐述超小世界性质的意义，我们来再看一下包含$N \approx 7\times 10^9$ 个节点的全球社会网络。假如社会是一个随机网络，那么节点间平均距离中对N 的依赖项是$\ln N$ =22.66。相反，如果该网络是无标度网络，那么对N 的依赖项为$\ln\ln N$ =3.12，这意味着枢纽节点可以大大缩短节点间的距离。

临界点状态（$\gamma=3$）

临界点状态有着特别的理论意义：γ 的值超过3之后，度分布的二阶矩就不再发散了。因此，人们将$\gamma=3$称为临界点（critical point）。在临界点状态下，随机网络中出现的平均距离对N的依赖$\ln N$又回来了。不过，计算中包含了一个双对数修正$\ln\ln N$ [29-31]。因此，和大小相同的随机网络相比，修正项使得临界点状态下的平均距离比随机网络的平均距离要短。

小世界（$\gamma>3$）

该状态下，$\langle k^2\rangle$是有限的，且平均距离遵循随机网络中出现的小世界性质。尽管枢纽节点依然存在，但在$\gamma>3$时，枢纽节点没有大到能够显著影响节点间平均距离的程度。

总结来说，公式**3.22**表明：枢纽节点越明显，就越能有效地缩短节点间的距离。图**3-12a**支持了这一结论，该图展示了不同的γ下无标度网络的平均路径长度。可以看出：尽管当N较小时，4种状态下的平均距离是相近的，但当N较大时，4种状态下的差异非常明显。

我们进一步研究了不同γ和N的无标度网络的路径长度分布（图**3-12b～d**）。当$N=10^2$时，不同γ所对应的路径长度分布几乎重叠，这表明此时γ对路径长度没有明显影响。但当$N=10^6$时，不同γ所对应的路径长度分布p_d被很好地区分开来了。图**3-12d**还表明，度指数越大，节点间的距离也越大。

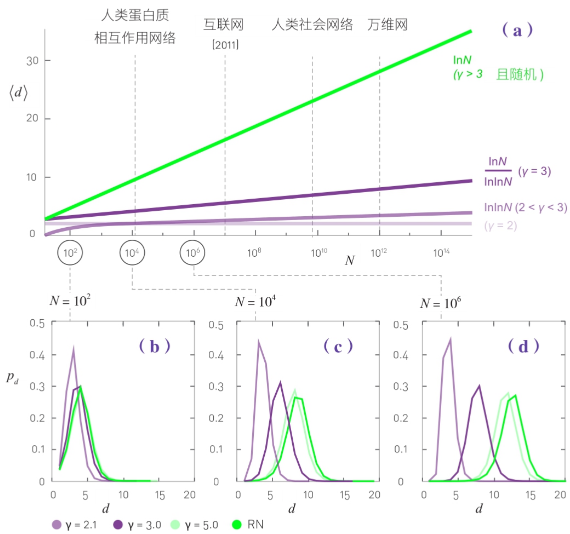


图3-12　无标度网络中的距离

（a）　该图描绘了无标度网络在4种不同状态下的平均路径长度。4种情况分别为：常数（$\gamma=2$），$\ln\ln N$（$2<\gamma<3$），$\ln N/\ln\ln N$（$\gamma=3$），$\ln N$（$\gamma>3$的无标度网络和随机网络）。虚线表示几个真实网络的近似大小。在生物网络中——例如人类的蛋白质相互作用网络（PPI），鉴于其规模不大，点到点的距离在4种不同状态下的差异相当小。而在人类社会网络和万维网这样大小的网络中，$\langle d\rangle$ 的差异非常显著，因此小世界的公式明显低估了这些网络的真实 $\langle d\rangle$。

（b）（c）（d）三幅图分别是大小为$N=10^2$，$N=10^4$，$N=10^6$的网络中的距离分布。该图说明，在小型网络（$N=10^2$）中，距离分布对γ不敏感。而对于大型网络（$N=10^6$）而言，p_d和$\langle d\rangle$会随着γ显著变化。

这些网络都是用静态模型[32]生成的，网络平均度为$\langle k\rangle=3$。

综上所述，无标度性质对网络距离有如下几点影响：

（1）　降低平均路径长度。因此，大多数具有实际意义的无标度网络不仅“小”，而且“超小”。这也是枢纽节点的存在带来的结果，枢纽节点在很多小度节点间扮演着桥梁的角色。

（2）　改变$\langle d\rangle$对系统大小的依赖关系，如公式**3.22**所示。度指数γ越小，节点间的距离越短。

（3）　只有在$\gamma>3$时，才会出现依赖关系$\ln N$——随机网络小世

界性质的体现（图**3-12**）。

边栏3.4

我们离枢纽节点总是很近

弗里杰什·考林蒂在他1929年的短篇小说[33]中首次描述了小世界概念，他提醒人们注意“在一群人中，我们很容易就能找出一个认识某位名人的人”。换句话说，我们离枢纽节点总是很近。这种效应在无标度网络中尤为明显（图**3-13**）。

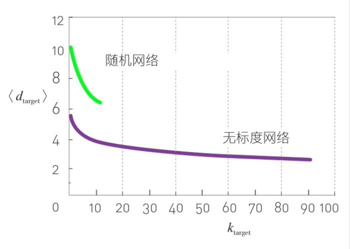


图**3-13**　和枢纽节点距离很近

图中分别描述了随机网络和无标度网络中度为$k \approx \langle k \rangle$的节点与度为$k_{target}$的目标节点之间的平均距离($d_{target}$)。和随机网络相比，无标度网络中其他节点距离枢纽节点更近。该图也说明了，在随机网络中最大度节点非常小，因此路径长度也比无标度网络中的明显更长。两个网络都有$\langle k \rangle = 2$，$N = 1\ 000$，无标度网络的度指数为$\gamma = 2.5$。

上述发现的意义很明显：我们和名人（例如著名科学家或美国总统）之间总是存在着很短的路径，这是因为这些人是拥有很多熟人的枢纽节点。这意味着，很多最短路径都要经过这些枢纽节点。

然而，和预期相反的是，我们在一项试图将六度概念复制到在线世界的实验中发现：和最终未能到达目标的路径中的节点相比，那些最终到达目标的路径中的节点更不愿意和枢纽节点联系。其中的原因其实是由个体自身造成的：人们认为枢纽节点非常忙，因此只有在真正需要的时候才会联系他们。在在线实验中，我们认为没必要联系枢纽节点，因此会尽量避开他们。

3.7 度指数的作用

无标度网络的许多性质都与度指数γ 有关。仔细查看表**3-1** 就会发现：

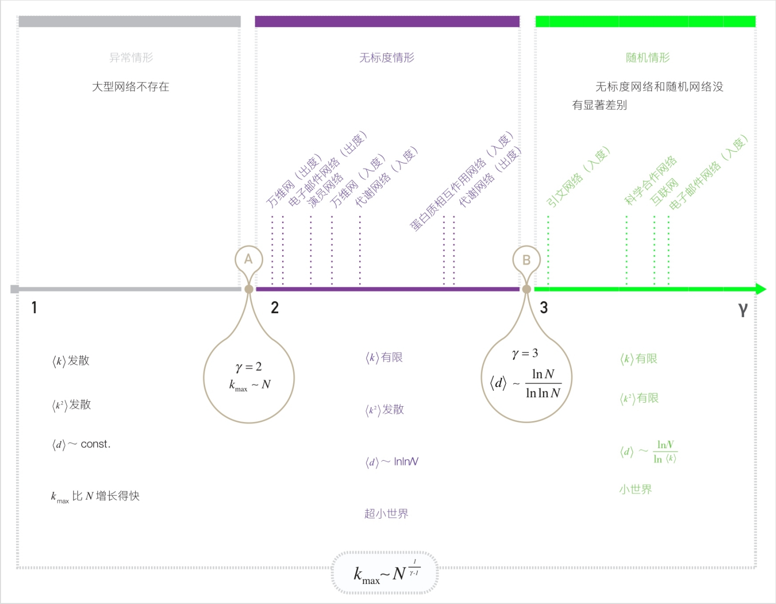
（1） 不同系统的γ 不相同，这促使我们研究网络性质是如何随着γ 变化的。

（2） 大多数真实系统的度指数都大于2，于是我们不禁会问：为什么没有看到γ ＜2的网络呢？

为了回答这些问题，接下来我们将讨论无标度网络的性质是如何随着γ 变化的（边栏**3.5** ）。

边栏3.5

无标度网络中和γ 相关的性质



异常情形（$\gamma<2$）

当$\gamma<2$时，公式**3.18** 中的指数$1/(\gamma-1)$比1大。因此，最大枢纽节点的链接数比网络大小增长得更快。这意味着当N 足够大时，最大枢纽节点的度一定会超过网络节点总数。也就是说，网络中所有的节点都不够用来连接最大枢纽节点。与之类似，当$\gamma<2$时，平均度$\langle k\rangle$在极限$N\to\infty$下发散。上述怪异的预测结果只是无标度网络在$\gamma<2$的异常情形下诸多怪异性质中的两个。这引发了一个更深层次的问题：在不允许多重链接存在的情况下，$\gamma<2$的大型无标度网络不存在（边栏**3.6** ）。

边栏3.6

为什么$\gamma<2$的无标度网络不存在

为了弄清楚为什么$\gamma<2$的网络不存在，让我们来尝试构建一个这样的网络。一个度序列如果可以转化为一个简单图（不包含多重链接和自环的图），则称该度序列是可图化的 [35] 。然而，并非所有的度序列都是可图化的：例如，如果度之和为奇数，则该序列不是可图化的（图**3-14b** ）。

度序列的可图化性可以使用埃尔德什和加莱（Gallai）提出的算法来判断 [35] ，[36] ，[37] ，[38] ，[39] 。如果将该算法应用到无标度网络中，我们会发现当$\gamma<2$时，可图化的度序列的数量为0（图**3-14c** ）。也就是说，任何一个$\gamma<2$的度分布都不能被转换成简单网络。实际上，处于该状态下的无标度网络，其枢纽节点的增长速度比N 快。如果不允许自环和多重链接出现，枢纽节点即使与所有节点相连，它的度也无法超过$N-1$。

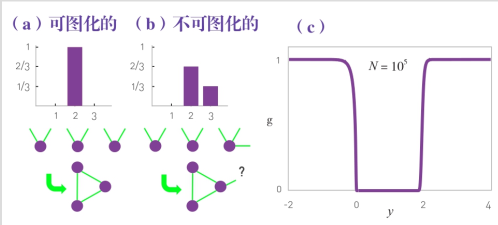


图3-14　γ ＜2的无标度网络不是可图化的

（a）（b）　是两个小型网络的度分布以及对应的度序列。两个度序列只有一个节点的度不同。使用度序列（图a）可以构造出一个简单网络，但无法用度序列（图b）来构造出一个简单网络——因为总有一个端无法匹配。因此，（a）是可图化的，而（b）是不可图化的。

（c）不同的γ 所对应的可图化的网络比例g 。给定γ ，我们生成了许多度指数为γ 、大小为N ＝10^5 的度序列，并检验每个度序列的可图化性。如图所示，虽然γ ＞2的网络几乎都是可图化的；但当0＜γ ＜2时，不存在可图化的无标度网络。[39]

无标度情形（2＜γ ＜3）

在这种情形下，度分布的一阶矩是有限的，但二阶矩及更高阶矩在极限$N \to \infty$下发散。因此，此时的无标度网络是超小的（3.6节）。公式3.18 预测，k_{max} 会随着网络大小的增长而增长，此时的度指数1/(γ -1)比1小。用k_{max}/N 表示最大枢纽节点的“市场占有率”——连接到它的节点占所有节点的比例，我们会发现该情形下最大枢纽节点的“市场占有率”会按照$k_{max}/N \sim N^{-(\gamma-2)/(\gamma-1)}$ 减少。

我们将在接下来的章节中看到，无标度网络的许多有趣特性——包括鲁棒性和异常传播现象，都是在这个情形下发生的。

随机情形（γ ＞3）

根据公式3.20 ，当γ ＞3时，度分布的一阶矩和二阶矩都是有限的。事实上，此时无标度网络的性质和同等大小的随机网络的性质并没有显著区别。例如，公式3.22 表明，节点间的平均距离会收敛于从随机网络中推导出的小世界公式。这是因为γ 较大时，度分布p_k 衰

减得非常快，枢纽节点的大小和数量都随之下降。

注意，当γ 较大时，无标度网络和随机网络没有显著区别。实际上，为了证明幂律度分布的存在，理想情况下我们需要将度分布跨越2到3个数量级。这意味着k_{max} 至少要比k_{min} 大10^2 ~10^3 倍。通过对公式**3.18** 求逆，我们可以估计出，要想观测到显著的无标度性质，所需要的网络大小为：

$$N=\left(\frac{k_{\max}}{k_{\min}}\right)^{\gamma-1} \quad (3.23)$$

例如，对于$\gamma=5$的无标度网络，如果要求其度分布跨越至少2个数量级（例如，$k_{min}\sim 1$，$k_{max}\simeq 10^2$），那么根据公式**3.23**，该网络的大小至少为$N>10^8$。实际上，很少有网络能够达到这样的规模。因此，虽然有些网络可能有大的度指数，但由于其规模有限，我们很难获得令人信服的证据来说明它们是无标度的。

综上所述，无标度网络的性质与度指数γ 密切相关。理论上讲，最有趣的情形是$2<\gamma<3$，此时$\langle k^2\rangle$ 发散，使得无标度网络超小。有趣的是，许多具有实际价值的网络，例如万维网和蛋白质相互作用网络，都对应着这个情形。

3.8 生成任意度分布的网络

通过埃尔德什 - 雷尼模型生成的网络具有泊松形式的度分布。然而，本章所讨论的实证结果表明，真实网络的度分布明显偏离于泊松形式。这带来了一个非常重要的问题：如何生成具有任意度分布p_k 的网络？本节我们将讨论解决该问题的三种常用算法。

配置模型

如图**3-15** 所描述的配置模型，可以帮助我们生成具有给定度序

列的网络。在该模型生成的网络中，每个节点具有预先定义好的度k_i，除此之外，网络是随机相连的。因此，这类网络通常被称为具有预定义度序列的随机网络。对于相同的度序列，按照配置模型生成网络的过程重复多次，我们可以得到具有相同度分布p_k 的不同网络（图**3-15b ～d** ）。这里，有几个需要特别注意的地方：

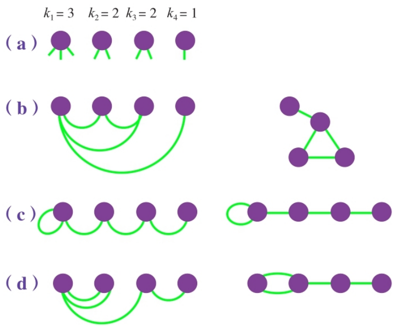


图3-15　配置模型

配置模型可以生成节点具有预定义度序列的网络[40]，[41]。算法包括如下两个步骤：

（**a**）度序列

度序列为每个节点赋予一个度，表示其拥有的“端”数或“半链接”数。度序列的产生有两种方式：一种是预定义的度分布p_k 中按照解析方式得到（边栏**3.7** ）；另一种是从真实网络的邻接矩阵中得到。开始时，端的个数必须是偶数，否则就会剩下不能匹配的端。

（**b**）～（**d**）网络生成

随机选择一对端进行连接。然后，在剩下的$2L-2$个端中再随机选取一对进行连接。重复该过程，直到所有的端都匹配成功。端的选择顺序不同，生成的网络也会不同。有些网络包含回路（图b），有些网络包含自环（图c）或多重链接（图d）。然而，在$N\to\infty$的极限情况下，自环和多重链接的期望数量趋近于0。

边栏**3.7**

生成符合幂律分布的度序列

无向网络的度序列是由节点的度构成的序列。例如，图**3-15a** 所示网络的度序列为{3, 2, 2, 1}。如图**3-15a** 所示，度序列不能唯

一确定一个图，这是因为同样的度序列可以按照不同的连接方式形成不同的图。

为了从预定义的度分布中生成一个度序列，让我们首先从一个按照解析方式定义出的度分布开始，例如$p_k \sim k^{-\gamma}$（如图**3-16a**所示）。我们的目标是生成一个服从分布p_k的度序列$\{k_1, k_2, \ldots, k_N\}$。我们首先计算函数

$$D(k) = \sum_{k' \geqslant k} p_{k'}, \tag{3.25}$$

如图**3-16b**所示。$D(k)$的值介于0和1之间，在k处的步长为p_k。为了生成一个服从度分布p_k且包含N个度的度序列，我们随机生成N个在(0,1)之间的随机数r_i，$i=1,\ldots,N$。对于每个r_i，我们使用子图b中的曲线得到对应的度k_i。按照这个方式得到的度$k_i = D^{-1}(r_i)$服从期望的度分布p_k。注意，同一个p_k所对应的度序列不是唯一的。也就是说，对于同一个p_k，我们可以生成多个度序列$\{k_1, k_2, \ldots, k_N\}$。

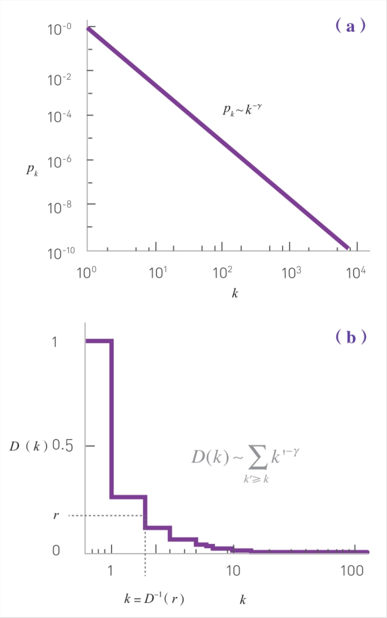


图3-16　生成度序列

（a）用于生成所期望的度序列的幂律度分布。

（b）公式**3.25** 使我们能够将度k 关联到均匀分布的随机数r 。

（**1**） 度为k_i 的节点和度为k_j 的节点之间有一条链接的概率为：

$$p_{ij} = \frac{k_i k_j}{2L-1} \tag{3.24}$$

注意，一条链接具有两个“端”，具有L 个链接的网络有$2L$ 个“端”。实际上，从节点i 发出的一个端可以连向其余$2L$ -1个端，而这些端中有k_j 个端是和节点j 相连的。因此，节点i 发出的一个端连向节点j 的概率是k_j /(2L -1)。由于节点i 有k_i 个端，因此它有k_i 次机会连向节点j ，从而得到公式**3.24** 。

（**2**） 由于算法没有禁止节点和自己相连或节点间形成多重链接，因此所生成的网络包含自环和多重链接。当然，我们也可以选择禁止上述情形出现，但如此一来，网络生成过程可能无法完成。拒绝自环或多重链接意味着，并不是所有可能配对的端都是等概率的。因此，公式**3.24** 将不再有效，使得分析计算变得困难起来。不过，可供选择的链接数随着N 增加，相比之下自环和多重链接的数量可以忽略不计，因此一般情况下，我们不需要排除自环和多重链接[42] 。

（**3**）配置模型在计算中经常被使用。这是因为公式**3.24** 和配置模型固有的随机性可以帮助我们对许多网络指标进行解析计算。

度保持的网络随机化

在探讨真实网络的性质时，我们经常会问：某个网络性质是否只用其度分布就可以预测出呢？或者说，该性质是否是度分布p_k 不能解释的新性质呢？为了回答这个问题，我们需要生成度分布p_k 和原网络相同但节点随机相连的网络。这可以通过图**3-17b** 所描述的度保持的

随机化（degree-preserving randomization）算法来实现。该算法背后的思想非常简单：随机选择两条链接，如果这两条链接交换之后不会产生多重链接，则进行交换。因此，链接交换涉及的4个节点的度在交换过程中不会改变。如此一来，枢纽节点依然是枢纽节点，度很小的节点度依然很小，但通过链接交换得到的网络变成随机连接的了。注意，度保持的网络随机化和完全随机化（full randomization）不同，完全随机化在链接交换过程中不保证节点的度不变（图**3-17a**）。实际上，完全随机化将任何网络都变成了埃尔德什-雷尼网络，最终得到的网络具有泊松度分布，该度分布和原网络的度分布p_k 是不相关的。

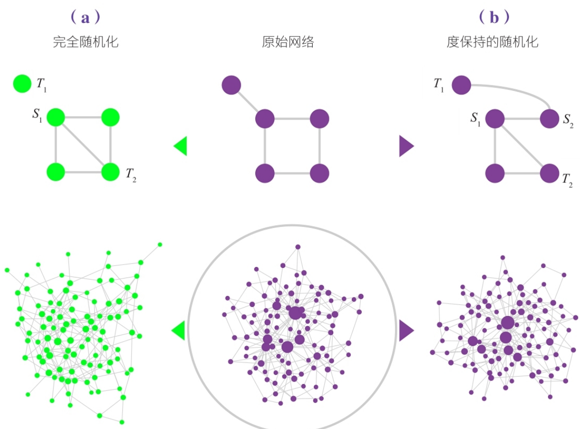


图**3-17**　度保持的随机化

有两种算法可以生成给定网络对应的随机化网络[43]，但这两种算法的结果不同。

（**a**）完全随机化

该算法生成的一个与原网络具有相同N 和L 的随机网络（埃尔德什-雷尼网络）。我们随机选择一个源节点（S_1）和两个目标节点，其中第一个目标节点（T_1）和源节点直接相连，而第二个目标节点（T_2）与源节点不直接相连。我们将链接S_1-T_1 断开，构造一个新链接S_1-T_2。这一过程中，目标节点T_1 和T_2 的度都改变了。对网络中的每条链接都进行一次上述操作，就会得到一个完全随机网络。

（**b**）度保持的随机化

该算法生成的网络中，每个节点的度都与原网络中对应节点的度相同，只是节点间的连接关系被随机化了。我们选择两个源节点(S_1，S_2)和两个目标节点(T_1，T_2)，初始时要求S_1 与T_1 相连，S_2 与T_2 相连。然后，我们将两条链接进行交换，使得S_1 与

T_2 相连，S_2 与T_1 相连。交换过程中，每个节点的度都保持不变。重复上述过程，直到每个链接都至少改变一次。

底部的三个子图分别是：原始的无标度网络（中间图）；完全随机网络（左图）——原始网络中的枢纽节点消失了；对原始网络进行度保持随机化得到的随机网络（右图）——保留了枢纽节点和原始网络的无标度性质。

隐参数模型

配置模型会生成自环和多重链接，而这些在真实网络中是不会出现的。我们可以用隐参数模型（**图3-18**）来生成具有预定义度分布p_k、且不含自环和多重链接的网络[44]，[45]，[46]。

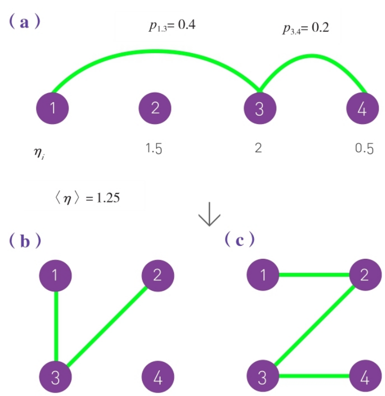


图3-18　隐参数模型

（**a**） 从N 个孤立的节点开始，为每个节点i 指定一个隐参数η_i。隐参数可以从预定义的分布ρ（η）中采样得到，也可以从给定的序列$\{\eta_i\}$中得到。每对节点相连的概率为：

$$p(\eta_i,\eta_j)=\frac{\eta_i\eta_j}{\langle\eta\rangle N}$$

图中给出了节点对(1,3)相连的概率以及节点对(3,4)相连的概率。

（**b**）（**c**） 节点之间按照连接概率完成连接之后，我们便得到了如图b和图c所示的网络，这两个网络是同样的隐参数模型（图a）生成的两个不同的网络。

该模型生成的网络中，期望链接数为：

$$L = \frac{1}{2}\sum_{i,j}^{N}\frac{\eta_i\eta_j}{\langle\eta\rangle N} = \frac{1}{2}\langle\eta\rangle N$$

类似于随机网络模型，链接数L 会随着网络的不同而改变，L 服从有界的指数分布。如果想控制平均度$\langle k\rangle$ ，我们可以将L 个链接逐个加入网络中。每个链接的两个端点i 和j 分别以概率η_i 和η_j 被随机选到。选到节点i 和j 之后，只有在它们之间还没有链接的情况下，我们才能将其连在一起。

我们从N 个孤立节点开始，为每个节点i 指定一个从分布$\rho(\eta)$中采样得到的隐参数η_i 。所生成网络的性质将会依赖于隐参数序列$\{\eta_i\}$的选择。有两种方式用于生成合适的隐参数：

- η_i 是从预定义好的分布$\rho(\eta)$中随机选取的N 个数字组成的序列。最终生成的网络，其度分布为：

$$p_k = \int \frac{e^{-\eta}\eta^k}{k!}\rho(\eta)d\eta \qquad (3.26)$$

- η_i 来自一个确定的序列$\{\eta_1,\eta_2,\ldots,\eta_N\}$。最终生成的网络，其度分布为：

$$p_k = \frac{1}{N}\sum_j \frac{e^{-\eta_j}\eta_j^{\ k}}{k!} \qquad (3.27)$$

隐参数模型提供了一种非常简单的用于生成无标度网络的方式。实际上，假如用

$$\eta_i = \frac{c}{i^\alpha}, i = 1,\cdots,N \qquad (3.28)$$

作为隐参数序列，那么根据公式**3.27** ，最终生成的网络在k 较大时具有如下形式的度分布：

$$p_k \sim k^{-\left(1+\frac{1}{\alpha}\right)} \qquad (3.29)$$

因此，我们可以通过选择合适的α 来调节度指数$\gamma=1+1/\alpha$ 。此外，由于公式**3.26** 和公式**3.27** 意味着$\langle k\rangle=\langle\eta\rangle$ ，我们也可以通过$\langle\eta\rangle$来调节$\langle k\rangle$。

综上所述，配置模型、度保持的网络随机化，以及隐参数模型都能生成具有预定义好的度分布的网络，从而可以帮助我们对网络的关键特性进行解析计算。每当要研究“某个特定网络性质是不是网络度分布的结果”或者“该网络性质是不是度分布之外网络中涌现出的一种新性质”时（边栏**3.8**），我们都会使用这几种算法来生成网络。在使用这些算法时，我们必须清楚它们的局限性：

- 这些算法并不能告诉我们，为什么网络具有某种特定的度分布。理解某个观测到的度分布p_k的形成原因，是第**5**章和第**6**章的主题。
- 一些重要的网络特性，包括聚集特性（第**8**章）和度相关性（第**6**章），在随机化过程中将会消失。

因此，这些算法生成的网络有点像一幅画的照片：乍一看和原画似乎是一样的，但仔细观察就会发现，包括画布质感和笔触在内的很多细节都丢失了。

前面讨论的三个算法还存在如下问题：如何确定使用哪一个呢？我们的选择将依赖于：我们是从一个度序列$\{k_i\}$开始的还是从一个度分布p_k开始的，以及我们是否能接受自环和多重链接的存在。图**3-19**给出了选择过程所涉及的决策树。

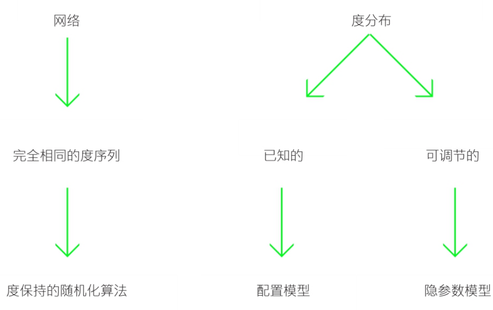


图**3-19**　生成算法选择

选择合适的生成算法取决于多个因素。如果我们从一个真实网络或已知的度序列开始，则可以选择度保持的随机化算法，该算法可以保证所生成的网络是简单网络（没有自环和多重链接），且具有和原始网络相同的度序列。

如果我们希望根据预定义的度分布p_k 来生成网络，则有两个选择。如果p_k 已知，配置模型提供了一种方便的网络生成算法。例如，配置模型可以生成一个具有纯幂律度分布网络，$p_k = Ck^{-\gamma}$ （$k \geq k_{min}$ ）。

然而，由于配置模型中唯一的自由参数是k_{min} ，调节无标度网络的平均度$\langle k \rangle$ 是一件非常烦琐的事情。因此，如果我们想调整$\langle k \rangle$ ，使用给定参数序列的隐参数模型（公式**3.28** ）会更方便。在这种方式下，度分布的尾部会服从幂律$\sim k^{-\gamma}$ ，而且可以通过改变链接数L 来控制$\langle k \rangle$。

边栏3.8

测试小世界性质

在文献中，真实网络中观察到的距离经常用来和小世界公式**2.19** 进行对比。然而，公式**2.19** 是从随机网络中推导出来的，而真实网络并不具有泊松度分布。如果网络是无标度的，则公式**3.22** 才是正确的公式。然而，公式**3.22** 只提供了大小为N 的网络中距离和网络大小N 的关系，并没有给出距离的绝对值。因此，除了拟合网络平均距离之外，我们经常要问的是：真实网络中观察到的距离，和相同度分布下的随机网络中观察到的距离是可比的吗？保持度序列的网络随机化方法可以帮助我们回答这个问题。下面，我们以蛋白质相互作用网络为例来阐述这一过程。

（**1**）原始网络

我们从测量原始网络中的距离分布p_d 开始，平均距离为$\langle d \rangle = 5.61$（图**3-20** ）。

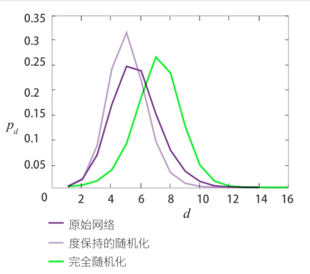


图**3-20**　真实网络的随机化

图中给出了蛋白质相互作用网络（表**3-1** ）中节点对间的距离分布p_d 。绿线表

示该网络在完全随机化后对应的路径长度分布。完全随机化使原始网络变成了一个埃尔德什 - 雷尼网络，只是N 和L 保持不变（图**3-17** ）。

浅紫色曲线对应的是度保持随机化之后的网络中的距离分布p_d 。随机化过程中，每个节点的度保持不变。

各个网络中节点平均距离为：$\langle d \rangle$ =5.61±1.64（原始网络），$\langle d \rangle$ =7.13±1.62（完全随机化后的网络），$\langle d \rangle$ =5.08±1.34（经过度保持的随机化得到的网络）。

（**2**）完全随机化

我们生成了一个和原始网络具有相同N 和L 的随机网络。所得到的距离分布p_d 明显右移了，平均距离$\langle d \rangle$ =7.13，比原始网络中的平均距离$\langle d \rangle$ =5.61大得多。人们不禁会得出这样的结论：蛋白质相互作用网络受到了某种未知的、能使距离变短的组织原则影响。但是，这个结论是有缺陷的：造成平均距离之间差异的原因可能是，完全随机化改变了网络的度分布。

（**3**）度保持的网络随机化

由于原始网络是无标度的，用于和原始网络进行对比的随机网络应该保持原始的度分布。因此，我们采用度保持的网络随机化方法生成了一个用于对比的随机网络，之后发现，所得到的距离分布p_d 和原始网络的p_d 是接近的。

综上所述，随机网络由于缺少枢纽节点而高估了节点间的距离。通过度保持的随机化得到的随机网络保留了枢纽节点，因此其距离和原始网络的距离非常接近。这个例子说明了，在研究网络时选择一个合适的随机化方法是非常重要的。

3.9 小结

无标度性质在网络科学的发展过程中发挥了非常重要的作用，主要有以下两点原因：

（1） 包括万维网和细胞网络在内的许多具有科学意义和实际意义的网络都是无标度的。这种普适性使得无标度性质成为许多学科都无法回避的问题。

（2） 一旦有枢纽节点存在，枢纽节点就会从根本上改变系统的行为。超小性就是枢纽节点对网络性质产生影响的例子之一。在接下来的章节里，我们将会看到更多的例子。

在进一步探索无标度性质的影响时，我们需要牢记：真实系统中很少会看到如公式**3.1** 那样纯粹的幂律分布形式。原因是，许多过程都会影响网络的拓扑结构，从而影响度分布的形状。在接下来的章节里，我们将讨论这些过程。这些过程的多样性以及度分布p_k 的复杂性，常常使那些仅从使用纯粹幂律分布对度分布进行拟合的角度来研究这些网络的人感到困惑。相反，无标度性质告诉我们，必须区分两种完全不同类型的网络：指数限界网络和重尾网络。

指数限界网络

如果一个网络的度分布在k 较大时成指数下降或以更快的速率下降，我们称该网络是指数限界的（exponentially bounded）。这类网络中，$\langle k^2\rangle$ 比$\langle k\rangle$ 要小，这意味着节点的度不存在显著差异。这类度分布p_k 包括泊松分布、高斯分布或简单的指数分布（表**3-2** ）。而埃尔德什-雷尼网络和瓦茨-斯托加茨网络是这一类别下最著名的两种网络模型。指数限界的网络中没有异常点，因此大多数节点的度相差不大。属于这一类别的真实网络有高速公路网络和电网。

重尾网络

如果一个网络的度分布在k 较大时服从幂律分布，我们称该网络为重尾的（fat tailed）。这类网络中，$\langle k^2\rangle$ 比$\langle k\rangle$ 大很多，节点的度差异很大。度分布服从幂律分布（公式**3.1** ）的无标度网络是这一类别下最著名的例子。异常点或者说度非常大的节点，在这些网络中不仅允许出现，更被期待出现。属于这一类别的真实网络包括万维网、互

联网、蛋白质相互作用网络，以及大多数社会网络。

尽管能够从统计上验证度分布的精确形式是最理想的，但一般来说，能够区分一个给定网络的度分布是指数限界的还是重尾的（进阶阅读**3.A**　）就已经足够了。如果度分布是指数限界的，则随机网络模型是理解其拓扑结构的合适起点。如果度分布是重尾的，则无标度网络是一个更好的近似。在接下来的章节中，我们将看到，重尾分布的关键特点是$\langle k^2 \rangle$　的数量级：如果$\langle k^2 \rangle$　较大，则系统表现得像一个无标度网络；如果$\langle k^2 \rangle$ 较小——和$\langle k \rangle$ 相近，则系统可以用随机网络很好地近似。

总之，要理解真实网络的性质，我们只需要记住：在无标度网络中，少数高度连接的枢纽节点和大量度很小的节点共存。这些枢纽节点的存在对于系统行为具有重要影响。在本章中，我们探索了无标度网络的一些基本特性，但也留下了一个重要的问题：为什么那么多真实网络是无标度的呢？下一章将给出答案。

边栏3.9

无标度网络一览

度分布

离散形式：

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

连续形式：

$$p\langle k \rangle = (\gamma - 1) k_{\min}^{\gamma-1} k^{-\gamma}$$

最大枢纽节点的大小

$$k_{\max} = k_{\min} N^{\frac{1}{\gamma-1}}$$

极限$N\to\infty$下，p_k 的矩

$2<\gamma\le 3$：$\langle k\rangle$是有限的，$\langle k^2\rangle$ 是发散的。

$\gamma>3$：$\langle k\rangle$和$\langle k^2\rangle$ 都是有限的。

距离

$$\langle d\rangle \sim \begin{cases} \text{const.} & \gamma=2 \\ \ln\ln N & 2<\gamma<3 \\ \dfrac{\ln N}{\ln\ln N} & \gamma=3 \\ \ln N & \gamma>3 \end{cases}$$

3.10 课后习题

习题1　枢纽节点

对于表3-1 列出的无向网络，计算其期望的最大度$k_{\max}$。

习题2　朋友悖论

度分布p_k 表示一个随机选择的节点拥有k 个邻居的概率。然而，如果我们随机选择一条链接，则其中一个端点的度恰好为k 的概率是$q_k=Akp_k$，这里A 是归一化因子。

（1）假设网络具有$2<\gamma<3$的幂律度分布，且最小度为$k_{\min}$、最大度为$k_{\max}$，计算出此时的归一化因子A。

（2）在配置模型中，q_k 也是一个随机选择的节点拥有一个度为k的邻居的概率。那么，对于一个随机选择的节点而言，其邻居的平均度是多少？

（3）假设一个网络满足$N=10^4$，$\gamma=2.3$，$k_{\min}=1$和$k_{\max}=1\,000$，那么随机选择一个节点，计算出其邻居的平均度，并

和该网络的平均度$\langle k\rangle$进行对比。

（4） 如何解释（3）中的“悖论”：一个节点的朋友总是比这个节点本身拥有更多的朋友？

习题3 生成无标度网络

编写计算机程序，生成大小为N、度分布服从度指数为γ的幂律分布的网络。具体过程参考**3.9**节。生成三个$\gamma=2.2$的网络，大小分别为$N=10^3$，$N=10^4$和$N=10^5$。每个网络中多重链接和自环所占的比例是多少？生成更多的网络，从而画出该比例随N变化的函数。设置$\gamma=3$，重复上述过程。

习题4 掌握分布

使用含有统计程序包的软件（例如Matlab、Mathematica和Python中的Numpy）生成3个人造数据集，每个数据集包含10 000个服从幂律分布的整数，3个数据集所对应幂律分布的指数分别为$\gamma=2.2$，$\gamma=2.5$和$\gamma=3$。生成数据集时，假设$k_{\min}=1$。采用进阶阅读**3.C**中介绍的技术来拟合这3个分布。

3.11 进阶阅读3.A 幂律分布

幂律分布在自然科学和社会科学领域有着复杂的历史。幂律分布也被交替地（有时是不正确地）称为胖尾分布、重尾分布、长尾分布、帕累托分布或布拉德福（Bradford）分布。幂律分布还有一系列的近亲，例如对数正态（log-normal）分布、韦伯（Weibull）分布和莱维（Lévy）分布。在本节中，我们将介绍网络科学中最常见的一些分布，并探讨这些分布与幂律分布的关系。

指数限界分布

自然界中的许多量，包括人类的身高和发生车祸的概率等，都服从限界分布。这类分布的共同特点是：当x 较大时，分布p_x 或者以指数（e^{-x} ）的速率衰减，或者以比指数更快的速率（ e^{-x^2/σ^2} ）衰减。因此，x 的期望最大值有一个上界$x_{\max}$ ，这个上界使得x 的期望最大值和$\langle x\rangle$ 相差不大。实际上，从限界分布p_x 中抽取N 个数时，我们会发现x 的期望最大值以$x_{\max}\sim\log N$ 或更慢的速度增长。这意味着，很大的x 值很少出现。实际上，很大的x 值非常罕见，其出现的概率几乎为零。同时，大多数从限界分布中抽取的x 值都在$\langle x\rangle$ 附近。

分布中x 较大的区域被称为分布的尾部。由于尾部出现的x 值非常少，因此限界分布有时也被称为瘦尾分布。

从解析角度讲，最简单的限界分布是指数分布$e^{-\lambda x}$ 。在网络科学中，最常遇到的限界分布是泊松分布（或二项分布）。泊松分布经常用于刻画随机网络的度分布。在网络科学之外，最常遇到的限界分布是正态（高斯）分布（表**3-2** ）。

重尾分布

术语胖尾、重尾和长尾是指在x 较大时衰减速率比指数慢的分布p_x 。在这些分布中，我们经常会遇到很大的x 值，这些值通常被称为异常点或罕见事件。公式**3.1** 所示的幂律分布，是重尾分布中最有名的例子。重尾分布的一个显著特征是，从该分布中抽取出的x 值可以跨越几个数量级。实际上，从这类分布中抽取x 值时，抽取N 次之后所得到的最大值可以达到$x_{\max}\sim N^{\zeta}$ ，其中ζ由刻画分布p_x 尾部的度指数γ 决定。由于N^{ζ} 增长很快，罕见事件或异常点出现的频率较大，而且通常主导着系统的性质。

重尾分布对于网络的意义体现在如下方面：

（1） 网络科学中的许多量，例如度、链接的权重、介数中心度等，在真实网络和模型网络中都服从幂律分布。

（2） 幂律形式可以通过合适的网络模型来分析预测（第**4**

章）。

交叉分布（对数正态分布，广延指数分布）

当实际观察到的分布看起来介于幂律分布和指数分布之间时，人们经常使用交叉分布（crossover distribution）对观测数据进行拟合。这类分布可能是指数限界的（具有指数截断的幂律分布），也可能是无限界的，但衰减速率比幂律分布（对数正态分布或广延指数分布）要快。接下来，我们将讨论几个常见交叉分布的性质。

具有指数截断的幂律分布，通常用于拟合真实网络的度分布，其密度函数具有如下形式：

$$p(x) = Cx^{-\gamma}e^{-\lambda x} \tag{3.30}$$

$$C = \frac{\lambda^{1-\gamma}}{\Gamma(1-\gamma, \lambda x_{\min})} \tag{3.31}$$

其中，$x > 0$且$y > 0$，$\Gamma(s, y)$表示上界不完全的伽马（gamma）函数。解析形式**3.30** 直接展现了这类分布的交叉特性：将重尾分布的关键部分——幂律项，和刻画指数限界尾部的指数项结合在一起。为了突出其交叉特性，我们对公式**3.30** 取对数，得到：

$$\ln p(x) = \ln C - \gamma \ln x - \lambda x \tag{3.32}$$

当$x \ll 1/\lambda$ 时，等式右边第二项起主要作用，意味着分布将服从指数为γ 的幂律分布。一旦$x \gg 1/\lambda$，λx 项将超过$\ln x$ 项的影响，从而导致在x 较大时，分布有一个指数截断。

广延指数分布（韦伯分布）在形式上类似于公式**3.30**，不同的是其指数项上有一个分数幂律。广延指数分布的名字源于这样一个事实：这类分布的累积分布函数等于1减去一个如公式**3.32** 的广延指数函数$P(x) = e^{-(\lambda x)^{\beta}}$，因此其密度函数为：

$$P'(x) = Cx^{\beta-1}e^{-(\lambda x)^{\beta}} \tag{3.33}$$

$$C = \beta\lambda^{\beta} \tag{3.34}$$

大多数应用中，x 介于0到$+\infty$之间。公式**3.33** 中，β 是广延指数，决定着$p(x)$的性质：

（**1**） 当$\beta=1$时，$p(x)$是一个简单的指数函数。

（**2**） 当β 介于0和1之间时，$\log p(x)$关于x 的曲线是“广延的”，即在x 轴上跨越了几个数量级。在这种情况下，很难将广延指数分布和纯幂律分布区分开来。β 越接近于0，$p(x)$就越接近幂律x^{-1}。

（**3**） 当$\beta>1$时，我们将得到一个“压缩的”指数函数，即x只在一个很小的范围内变化。

（**4**） 当$\beta=2$时，公式**3.33** 退化为瑞利分布（Rayleigh distribution）。

我们在第**4**章 和第**5**章 中将会看到，一些网络模型生成的网络具有广延指数形式的度分布。

如果$\ln x$ 服从正态分布，我们说x 服从对数正态分布（也称高尔顿或吉布拉分布）。一般而言，服从对数正态分布的变量是许多正的独立随机变量的乘积。例如，我们在金融领域遇到的对数正态分布表示一系列交易的综合收益。

对数正态分布的概率密度函数为：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right] \quad (3.35)$$

对数正态分布和正态分布很像，不同的地方是，它的指数项中是$\ln x$ 而非x。

为了理解为什么对数正态分布有时会被用来拟合幂律分布，我们要注意：

$$\sigma^2=\left\langle(\ln x)^2\right\rangle-\langle\ln x\rangle^2 \quad (3.36)$$

体现了x 在数量级上的典型变化幅度。由于$\ln x$ 服从正态分布，意味着x 的变化范围可以非常大。依赖于σ 的取值，对数正态分布在几个

数量级上可以和幂律分布非常接近。如表**3-2** 中所示，$\langle x^2\rangle$ 关于σ 成指数级增长，因此$\langle x^2\rangle$ 可以非常大。

表3-2　网络科学中的分布

名称	$p_x/p(x)$	$\langle x\rangle$	$\langle x^2\rangle$
泊松分布 （离散）	$\mathrm{e}^{-\mu}\mu^x/x!$	μ	$\mu(1+\mu)$
指数分布 （离散）	$(1-\mathrm{e}^{-\lambda})\mathrm{e}^{-\lambda x}$	$1/(\mathrm{e}^{\lambda}-1)$	$(\mathrm{e}^{\lambda}+1)/(\mathrm{e}^{\lambda}-1)^2$
指数分布 （连续）	$\lambda\mathrm{e}^{-\lambda x}$	$1/\lambda$	$2/\lambda^2$
幂律分布 （离散）	$x^{-\alpha}/\zeta(\alpha)$	$\begin{cases}\zeta(\alpha-2)/\zeta(\alpha), & \text{if } \alpha>2\\ \infty, & \text{if } \alpha\leqslant 1\end{cases}$	$\begin{cases}\zeta(\alpha-1)/\zeta(\alpha), & \text{if } \alpha>1\\ \infty, & \text{if } \alpha\leqslant 2\end{cases}$
幂律分布 （连续）	$\alpha x^{-\alpha}$	$\begin{cases}\alpha/(\alpha-1), & \text{if } \alpha>2\\ \infty, & \text{if } \alpha\leqslant 1\end{cases}$	$\begin{cases}\alpha/(\alpha-2), & \text{if } \alpha>1\\ \infty, & \text{if } \alpha\leqslant 2\end{cases}$
带指数截断的幂律分布 （连续）	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha)}x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\lambda^{-1}\frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}$	$\lambda^{-2}\frac{\Gamma(3-\alpha)}{\Gamma(1-\alpha)}$
广延指数分布 （连续）	$\beta\lambda^{\beta}x^{\beta-1}\mathrm{e}^{-(\lambda x)^{\beta}}$	$\lambda^{-1}\Gamma(1+\beta^{-1})$	$\lambda^{-2}\Gamma(1+2\beta^{-1})$
对数正态分布 （连续）	$\frac{1}{x\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(\ln x-\mu)^2/(2\sigma^2)}$	$\mathrm{e}^{\mu+\sigma^2/2}$	$\mathrm{e}^{2(\mu+\sigma^2)}$
正态分布 （连续）	$\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$	μ	$\mu^2+\sigma^2$

表中列出了网络科学中常见的一些分布。对于每个分布，我们给出了其密度函数p_x。这里，我们选择合适的归一化常数C ，使其满足：

连续情况

$$\int_{x=x_{\min}}^{\infty} Cf(x)dx=1$$

离散情况

$$\sum_{x=x_{\min}}^{\infty} Cf(x)=1$$

由于$\langle x\rangle$ 和$\langle x^2\rangle$ 在网络科学中起着重要的作用，因此对于每个分布，我们在表中都给出了这两个量的解析形式。有些分布在$x=0$处发散，因此在计算$\langle x\rangle$ 和$\langle x^2\rangle$ 的过程

中，我们假设系统中存在一个小的截断x_{min}。在网络中，x_{min}通常对应着最小度k_{min}或者是能够保证最佳拟合的最小度。

总之，在重尾分布出现的大多数领域中，一直存在着这样的争论：到底哪种分布可以对这类数据做出最佳拟合。常见的候选分布包括：幂律分布、广延指数分布、对数正态分布。在很多系统中，经验数据并不足以区分这些分布。因此，只要有经验数据需要被拟合，围绕着最佳拟合的争论就不会停止。

这场争论最终靠准确的机理模型来解决——这些模型能解析并预测出期望的度分布。我们在接下来的章节中将看到，在网络背景下，这些模型能预测出泊松分布、简单指数分布、广延指数分布和幂律分布。表**3-2**　中的其余分布有时也会被用来拟合一些网络的度分布——尽管我们缺少判断这些分布是否和网络相关的理论基础。

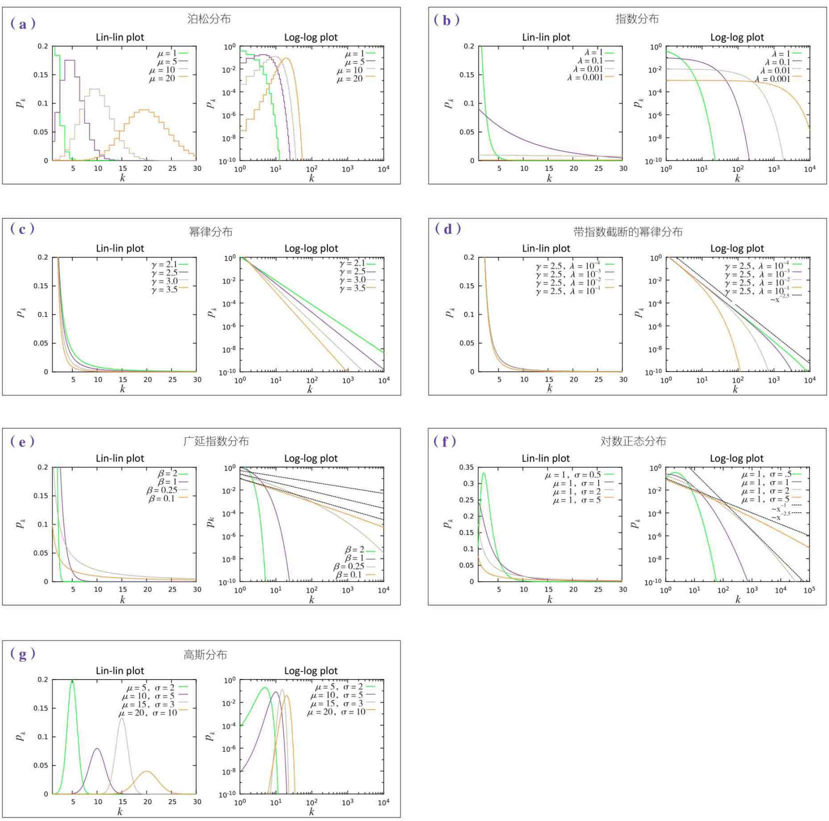


图3-21　分布可视化

线性和双对数坐标下绘制出的网络科学中最常见的一些分布。关于这些分布的定义，可以参见表3-2 。

3.12
进阶阅读3.B
绘制幂律分布

绘制度分布是分析网络性质时必不可少的一部分。绘制度分布之前，我们首先需要得到度为k 的节点数N_k 。N_k 可以通过直接测量得到，也可以从网络模型中计算得到。有了N_k ，我们可以计算出度分布$p_k = N_k / N$ 。问题是，如何绘制p_k 才能最好地体现出度分布的性质呢？

使用双对数坐标

在无标度网络中，大量只有一两个链接的节点和少量具有数千甚至数百万个链接的枢纽节点共同存在。如果我们使用线性横坐标轴，则包含大量节点的小k 区域会因被压缩而不明显。与之类似，由于p_k在$k = 1$和k 较大时的取值会相差几个数量级，使用线性纵坐标会使k较大时的p_k 值看起来像零（图**3-22a** ）。双对数坐标可以避免上述问题。我们可以使用以10为底的对数坐标轴（本书中使用了这种方式，如图**3-22b** ），也可以在线性坐标下绘制$\log p_k$ 关于$\log k$ 的函数（同样正确，但图会比较难理解）。注意，由于$\log 0 = -\infty$ ，在双对数坐标下，$p_k = 0$和$k = 0$对应的点无法显示。

避免线性分箱

最有缺陷的方法（却经常在文献中看到）是简单地将$p_k = N_k / N$绘制在双对数坐标下（图**3-22b** ）。这种方法被称为线性分箱（linear binning），每个分箱的大小都一样：$\Delta k = 1$。对于无标度网

络，线性分箱会在k 较大时形成一个明显的平缓区，即大量数据点构成一条水平线（图3-22b ）。

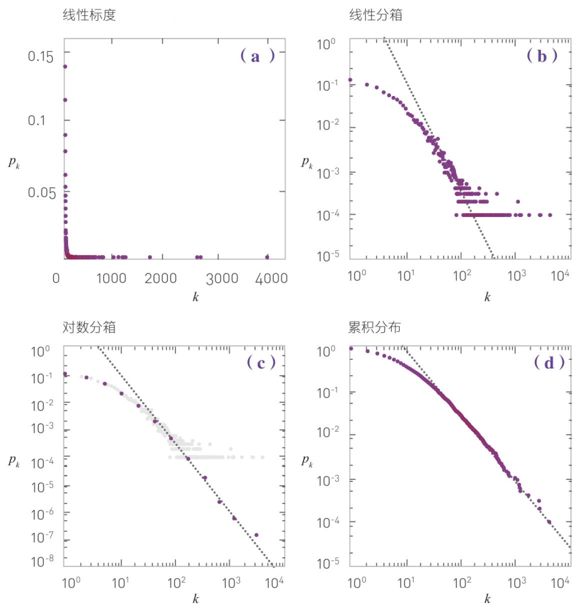


图3-22　绘制度分布

使用文中描述的4种方法绘制的度分布$p_k \sim (k+k_0)^{-\gamma}$，其中$k_0=10$，$\gamma=2.5$。

（**a**）线性坐标，线性分箱

在线性坐标下，幂律分布很难看清。因此，我们通常会使用双对数坐标来绘制无标度网络的度分布。

（**b**）双对数坐标，线性分箱

此时，分布的尾部清晰可见，但在k 较大的区域有一个平缓区，这是线性分箱的结果。

（**c**）双对数坐标，对数分箱

对数分箱使得平缓区消失了，标度扩展到了k 较大的区域。为便于对照，我们用浅灰色给出了（b）中线性分箱的结果。

（**d**）双对数坐标，累积分布

双对数坐标下的累积度分布。

平缓区的出现有一个简单的解释：通常情况下，k 很大时，度相

同的节点往往只有一个；因此，在k 较大的区域，要么N_k $=0$（没有度为k 的节点），要么N_k $=1$（只有一个度为k 的节点）。如此一来，线性分箱会使得p_k $=0$（在双对数坐标下无法显示）或p_k $=1/N$（适用于所有枢纽节点，在p_k $=1/N$ 处形成一条水平线）。

平缓区的出现会影响我们对于度指数γ的估计。例如，使用线性分箱法来拟合图**3-22b** 中服从幂律分布的数据时，得到的γ和真实值γ $=2.5$差异非常大。原因是：在线性分箱下，k 较小的区域有大量的节点，使我们可以很好地拟合p_k ；k 较大的区域中节点非常少，无法对p_k 进行准确统计。也就是说，平缓区的出现使我们对p_k 的估计有偏差。而实际上，恰恰是k 较大的区域在对γ 的估计中起着决定性作用。增大分箱的大小也无法解决这个问题。因此，在拟合重尾分布时应避免使用线性分箱法。

使用对数分箱

对数分箱修正了线性分箱中的非均匀采样问题。在对数分箱中，分箱的大小随着度的增大而增大，从而确保每个分箱中的节点数大致相当。例如，分箱大小可以设置为2的倍数：第一个分箱的大小为b_0 $=1$，由所有度k $=1$的节点构成；第二个分箱的大小为b_1 $=2$，由所有度k $=2,3$的节点构成；第三个分箱的大小为b_2 $=4$，由所有度k $=4,5,6,7$的节点构成。归纳起来，第n 个分箱的大小为2^{n-1} ，由所有度k $=2^{n-1}, 2^{n-1}+1, \ldots, 2^n-1$的节点构成。注意，分箱大小的增幅可以任意设置，即$b_n$ $=c^n$ （c >1）。在对数分箱中，度分布为$p_{\langle k_n\rangle}=N_n/b_n$ ，其中N_n 是大小为b_n 的第n 个分箱中所有节点的个数，$\langle k_n\rangle$ 是分箱b_n 中所有节点的平均度。

图**3-22c** 展示了对数分箱下的p_k 。注意，此时标度扩展到了k 较大的区域，而这个区域在线性分箱下是不可见的。因此，对数分箱能从数量稀少的大度节点中提取出有用信息（边栏**3.10** ）。

使用累积分布

另一种从p_k 的尾部提取信息的方式是绘制互补累积分布：

$$P_k = \sum_{q=k+1}^{\infty} p_q \tag{3.37}$$

这种方式同样增强了度较高区域的统计显著性。如果p_k 服从幂律分布（公式**3.1** ），那么累积分布的标度为：

$$P_k \sim k^{-\gamma+1} \tag{3.38}$$

累积分布消除了线性分箱中出现的平缓区，从而得到一个延伸了的标度区域（图**3-22d** ），使我们能够对度指数进行更精确的估计。

总之，绘制度分布以展现其特点是一件需要特别注意的事。掌握合适的工具可以帮助我们更好地探索真实网络的性质（边栏**3.10** ）。

边栏**3.10**

真实网络的度分布

在真实系统中，我们很少看到服从纯粹幂律的度分布。实际上，大多数真实系统的p_k 具有如图**3-23a** 所示的形状，呈现出一些经常出现的特点：

（**1**） 低度饱和（Low- degree saturation）是一种常见的偏离幂律分布的情况。低度饱和表现为：$k < k_{sat}$ 时p_k 是平的。这表明，小度节点的数量比纯幂律分布所预期的要少。我们将在第**5**章 中解释低度饱和出现的原因。

（**2**） 高度截断（High- degree cutoff）是指度分布p_k 在$k > k_{cut}$ 时有一个快速下降。这表明，大度节点的数量比纯幂律分布所预期的要少。这限制了最大枢纽节点的大小，使其比公式**3.18** 所预测的要小。如果节点可以拥有的链接数存在固有的限制，高度截断就会发生。例如，在社交网络中，一个人很难和大量的人都维持着熟人关系。

鉴于这种截断的广泛存在，度分布有时会用下面的式子来拟合：

$$p_k = a(k + k_{sat})^{-\gamma} \exp\left(-\frac{k}{k_{cut}}\right) \quad (3.39)$$

其中，k_{sat} 用于解释低度饱和，而指数项用于解释高度截断。为了展示标度的完整范围，我们通常使用下面的公式：

$$\tilde{p}_k = p_k \exp\left(\frac{k}{k_{cut}}\right) \quad (3.40)$$

令$\tilde{k} = k + k_{sat}$，根据公式3.40，修正了两种截断之后，有$\tilde{p} \sim \tilde{k}^{-\gamma}$如图3-23b 所示。

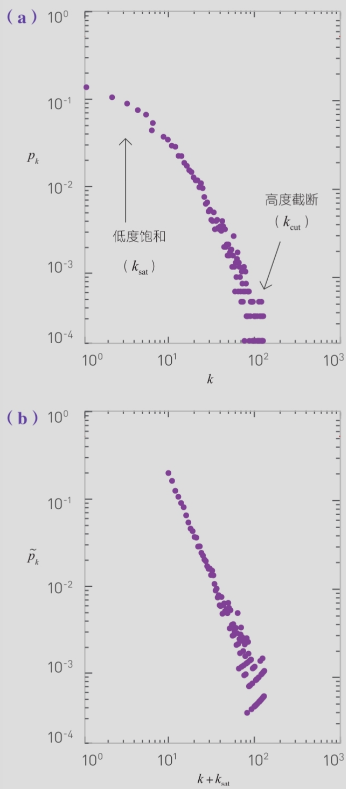


图3-23　标度变换后的度分布

（a）在真实网络中，由于存在低度饱和和高度截断，度分布经常偏离纯幂律分布。

（b）按照公式3.40 进行标度变换后，我们绘制出$\tilde{p}_k$ 随着($k + k_{sat}$)变化的函

数。可以看出，度分布对于所有的度都服从幂律分布。

人们有时会认为，低度和高度截断的出现意味着网络不是无标度的。这是对无标度性质的一种误解：实际上，无标度网络的所有性质对低度饱和都是不敏感的。只有高度截断会影响系统的性质，因为高度截断会限制二阶矩$\langle k^2 \rangle$ 的发散。这些截断的出现表明，还有一些其他的现象需要我们去理解。

3.13 进阶阅读3.C 估计度指数

由于无标度网络的性质依赖于度指数（3.7节 ），因此我们需要确定度指数γ 的值。然而，在试图用幂律分布来拟合真实数据的时候，我们遇到了一些困难。最重要的困难是，度指数所代表的标度并非在度分布的整个取值范围内都有效。相反，我们会看到，度分布存在低度和高度截断（边栏3.10 ）——在本节中分别用$K_{\min}$和$K_{\max}$表示，而在二者之间的区域，我们可以看到一个清晰的标度。注意，$K_{\min}$和$K_{\max}$不同于表示网络中节点最小度和最大度的$k_{\min}$和$k_{\max}$。$K_{\min}$和$K_{\max}$与边栏3.10 中讨论的k_{sat}和k_{cut}相同。这里，我们主要讨论如何估计低度截断$K_{\min}$，而高度截断可以用类似的方法确定。建议读者在实践这些方法之前，先查阅一下本节最后给出的关于系统问题的一些讨论。

拟合方法

由于度分布通常表示为一系列正整数$k_{\min},\ldots,k_{\max}$，我们主要考虑如何从一组离散数据点中估计γ [47] 。我们使用引文网络来说明这一过程。网络中包括$N=384\ 362$个节点，每个节点代表一篇1890年至2009年间发表在美国物理学会期刊上的学术论文。网络中包括$L=2\ 353\ 984$条链接，每条链接表示一篇论文对数据集中其他论文的引用

（此处不考虑数据集之外的引用）。这个引文网络并不是表**3-1** 中所列举的引文网络。关于这个引文网络数据集的完整描述，请参见文献[48]。拟合过程的步骤如下[47]：

在线资源**3.2**

合适的幂律

执行本节中描述的拟合过程的算法工具可在http://tuvalu.santafe.edu/~aaronc/Powerlaws/上找到。

扫描二维码下载“湛庐阅读”App，回复书名，获取算法工具。

（**1**） 在$k_{\min}$和$k_{\max}$之间选择一个合适的$K_{\min}$。根据下面的式子，估计出此$K_{\min}$所对应的度指数

$$\gamma = 1 + N\left[\sum_{i=1}^{N}\ln\frac{k_i}{K_{\min}-\frac{1}{2}}\right]^{-1} \qquad (3.41)$$

（**2**）根据得到的参数对(γ，$K_{\min}$)，并假设度分布有如下形式：

$$p_k = \frac{1}{\zeta(\gamma, K_{\min})}k^{-\gamma} \qquad (3.42)$$

则对应的累积分布函数（CDF）为

$$P_k = 1 - \frac{\zeta(\gamma, k)}{\zeta(\gamma, K_{\min})} \qquad (3.43)$$

（**3**） 使用柯尔莫哥洛夫 - 斯米尔诺夫检验（Kormogorov-Smirnov test）来确定数据$S(k)$的CDF和公式**3.43** 在参数对(γ，$K_{\min}$)下拟合得到的累积分布之间的最大距离D：

$$D = max_{k \geqslant K_{min}} \left| S(k) - P_k \right| \quad (3.44)$$

公式3.44 可以得到使经验分布$S(k)$和拟合分布（公式3.43 ）间的差异D 最大的度。

（4） 对k_{min} 到k_{max} 范围内的所有K_{min} ，重复步骤（1）～（3）。我们的目标是找到使公式3.44 中的D 值最小的K_{min}。为了说明这一过程，我们以引文网络为例画出了D 关于K_{min} 的函数（图3-24b ）。如图所示，当$K_{min}=49$时，D 取到最小值。此时，由公式3.41 得到的代表着最佳拟合的γ 取值为$\gamma=2.79$。所得到的度指数的标准差为：

$$\sigma_{\gamma} = \frac{1}{\sqrt{N\left[\frac{\zeta''(\gamma, K_{min})}{\zeta(\gamma, K_{min})} - \left(\frac{\zeta'(\gamma, K_{min})}{\zeta(\gamma, K_{min})}\right)^2\right]}} \quad (3.45)$$

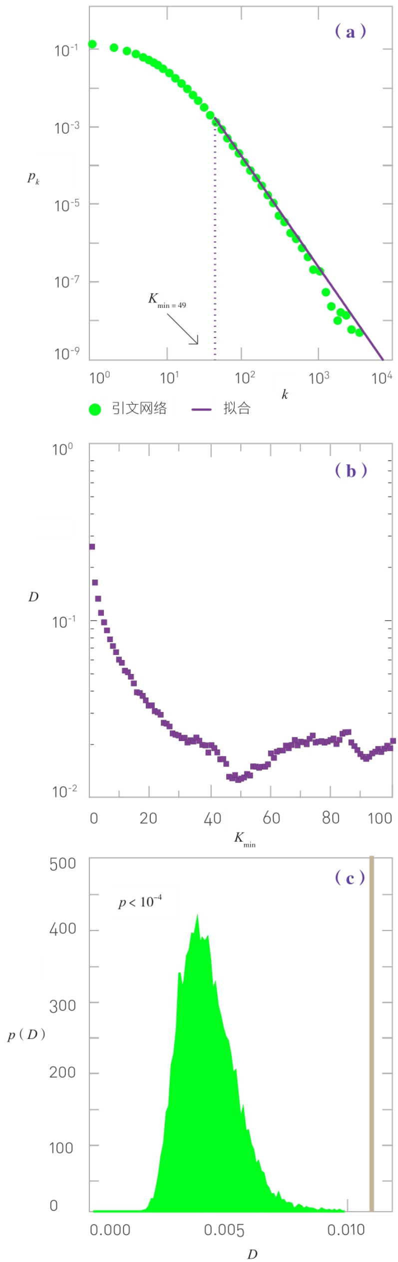


图3-24　最大似然估计

（**a**）引文网络的度分布p_k，紫色直线表示基于模型（公式**3.39**）的最佳拟合。

（**b**）引文网络中，柯尔莫哥洛夫－斯米尔诺夫检验值关于K_{min}的变化曲线。

（**c**）M ＝10 000个生成数据集上得到的p（$D_{synthetic}$），灰色线对应的是从引文网络中得到的D_{real}值。

这意味着最佳拟合是$\gamma \pm \sigma_{\gamma}$。对于该引文网络而言，我们有$\sigma_{\gamma}$＝0.003，因此$\gamma$＝2.79(3)。

注意，为了估计γ，小于N＝50的数据集应小心使用。

拟合优度

得到了使用幂律分布拟合数据集的最佳参数对(γ，K_{min})，并不意味着幂律分布本身是拟合该数据集的一个好模型。因此，我们需要通过拟合优度进行检验。拟合优度产生一个p值，用来量化幂律分布假设的合理性。最常用的方法包括以下步骤：

（**1**）用累积分布（公式**3.43**）来估计真实数据和最佳拟合之间的KS距离，用D_{real}表示。这对应着前面拟合过程的第三步，我们从中得到了数据最佳拟合K_{min}所对应的D值。对于引文数据，最佳拟合为K_{min}＝49，此时有D_{real}＝0.011 58（图**3-24c**）。

（**2**）根据公式**3.42**生成包含N个度的度序列（随机数的个数和网络中的节点数相同），并用生成的度序列代替真实数据，得到这个生成的度序列所对应的$D_{synthetic}$。因此，$D_{synthetic}$代表了和我们的度分布一致的生成度序列同真实数据之间的距离。

（**3**）我们的目标是检查D_{real}和$D_{synthetic}$是否相近。为此，我们重复M次（$M \gg 1$）步骤（2），每次生成一个新的度序列并确定其所对应的$D_{synthetic}$，从而得到一个分布p（$D_{synthetic}$）。我们绘制出p（$D_{synthetic}$），并用竖线标记出D_{real}出现的位置（图**3-24c**）。如果D_{real}在p（$D_{synthetic}$）分布之内，则意味着：最佳拟合所对应的分布和真实数据之间的距离，与从最佳拟合分布中采样得到的度序列和真实数据之间的距离相比

是接近的。因此，幂律分布是该数据的一个合理模型。然而，如果D^{real}在分布$p(D^{synthetic})$之外，则幂律分布不是该数据的一个好模型——存在其他函数形式的分布，能更好地刻画原始分布p_k。

如图**3-24c** 所示的分布有时候已经足以说明拟合的统计显著性了。但一般而言，最好能够给出一个具体的值来刻画一个分布的合适性，即：

$$p = \int_{D^{real}}^{\infty} P\left(D^{synthetic}\right) dD^{synthetic} \tag{3.46}$$

p 值越接近1，越能说明经验数据和模型之间的距离可以归因为统计扰动。如果p 很小，则说明该模型不适合用于拟合该数据。

一般而言，如果$p > 1\%$，模型可以被接受。对于引文网络，使用幂律分布进行拟合时有$p < 10^{-4}$，这说明纯幂律分布并不是原始度分布的合适模型。这个结果多少有些令人惊讶，毕竟引文数据的幂律分布性质自20世纪60年代以来就被反复报道[7]，[8]。幂律分布对引文网络度分布拟合的失效说明，在对分布缺少解析理解的情况下盲目套用幂律分布进行拟合是有局限性的。

拟合真实分布

为了解决这一问题，我们注意到公式**3.44** 在拟合数据时排除了所有$k < K_{min}$的数据点。由于引文网络是重尾的，$K_{min} = 49$使得超过96%的数据点被排除在外了。实际上，区域$k < K_{min}$中包含了统计上有用的信息，而这些信息在前面的拟合中被忽略了。我们需要引进一种替代的模型来解决这个问题。

正如我们在边栏**3.10** 中讨论的那样，许多像引文网络这样的真实网络，其度分布并不服从纯幂律分布。真实网络的度分布通常有低度饱和和高度截断，可以描述为如下形式：

$$p_k = \frac{1}{\sum_{k'=1} (k' + k_{sat})^{-\gamma} e^{-k'/k_{cut}}} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}} \tag{3.47}$$

对应的CDF为：

$$P_k = \frac{1}{\sum_{k'=1}(k' + k_{\text{sat}})^{-\gamma} e^{-k'/k_{\text{cut}}}} \sum_{k'=1}^{k} (k' + k_{\text{sat}})^{-\gamma} e^{-k'/k_{\text{cut}}} \quad (3.48)$$

其中，k_{sat} 和k_{cut} 分别对应低度饱和和高度截断。与之前的方法的不同之处在于，我们现在没有舍弃偏离纯幂律分布的数据点，而是选择了对度分布整体（从$k_{\min}$到$k_{\max}$）拟合效果更佳的函数。

我们的目标是找到模型（公式**3.47**）的拟合参数k_{sat}、k_{cut}和γ，步骤如下（图**3-25**）：

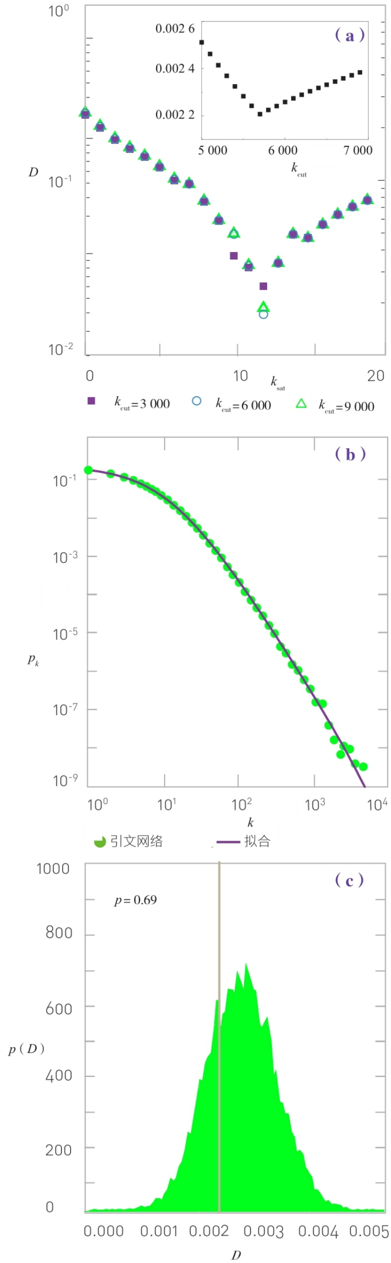


图3-25　估计引文网络的标度参数

（a） 柯尔莫哥洛夫 - 斯米尔诺夫参数D 关于k_{sat} 的变化曲线，三条线分别对应着k

$_{cut}$ = 3 000,6 000,9 000。图中曲线表明，当k_{sat} = 12时，对应的D 取到最小值。子图：k_{sat} = 12时，D 关于k_{cut} 的变化曲线，D 在k_{cut} = 5 691处取值最小。

（b） 度分布p_k ，其中直线表示由（a）得到的最佳估计。此时，模型正确地拟合了整条曲线，不仅包括曲线的尾部，也不仅是像图**3-24a** 中所做到的那样。

（c） M = 10 000个生成数据集上得到的$p(D^{synthetic})$，灰色线对应的是从引文网络中得到的D^{real} 值。

（1） 在k_{min} 到k_{max} 之间分别为k_{sat} 和k_{cut} 选择一个值。使用梯度下降法最大化对数似然函数

$$\log\mathcal{L}(\gamma \mid k_{sat},k_{cut}) = \sum_{i=1}^{N}\log p(k_i \mid \gamma,k_{sat},k_{cut}) \qquad (3.49)$$

估计出度指数γ 。也就是说，固定(k_{sat}，k_{cut})，找到使公式**3.49** 取最大值的γ 。

（2） 假设度分布服从公式**3.47** 的形式，参数为上一步得到的γ (k_{sat}，k_{cut})。计算出原始数据的累积度分布（CDF）和由公式**3.47** 得到的拟合模型之间的柯尔莫哥洛夫 - 斯米尔诺夫参数D 。

（3） 从k_{min} = k_0 到k_{max} 遍历k_{sat} ，从k_{min} = k_0 到k_{max} 遍历k_{cut} ，重复步骤（1）～（3）。目标是找到使D 最小的k_{sat} 和k_{cut} 。为了说明该过程，使用引文网络数据，我们在图**3-25a** 中绘制了给定k_{cut} 时D 关于k_{sat} 的函数。使D 最小的(k_{sat}，k_{cut})以及由公式**3.41** 得到的γ ，就是最终的最佳拟合参数。在我们的引文网络数据集中，最佳拟合参数为k_{sat} = 12，k_{cut} = 5691，γ = 3.028。我们发现，真实数据的D 在生成的分布$p(D)$之内（图**3-25c** ），对应的p 值为69%。

拟合问题总结

前面描述的方法也许会让你觉得，确定度指数是一个烦琐但直观的过程。实际上，这些拟合方法有一些众所周知的局限：

（1） 纯幂律分布是一个理想的分布，公式**3.1** 这样的形式只在一

些简单的模型中会出现（第4章 ）。现实中，很多过程都会影响真实网络的拓扑结构，从而影响度分布的形状。我们将在第5章 中讨论这些过程。如果p_k不服从纯幂律分布，则前面介绍的对数据进行幂律拟合的方法势必不能检测到统计显著性。虽然这可能意味着网络不是无标度的，但通常也意味着我们对度分布的准确形式缺少正确的认识。因此，我们可能在用错误的函数形式对数据集的p_k进行拟合。

（2） 前述用于检验拟合优度的统计工具依赖于柯尔莫哥洛夫 - 斯米尔诺夫标准——度量拟合模型和数据集之间的最大距离。如果几乎所有的数据点都服从完美的幂律分布，但有一个数据点因为某种原因偏离了曲线，我们也会因此而失去拟合的统计显著性。在真实系统中，有很多原因会造成这种局部偏离，而这种偏离对系统的整体行为影响甚微。然而，去掉这些“异常点”会被看作“数据操纵”；如果保留这些点，我们则无法检测到幂律分布拟合的统计显著性。

一个好例子是演员网络，其度分布在大多数度上都符合幂律分布。但是，在$k=1\ 287$处有一个明显的异常点——源于1956年的一部电影：《环游世界八十天》。这部电影是imdb.com（演员网络的数据来源）上唯一一部列出了所有临时演员的电影，这部电影有1 288位演员。而演员人数第二多的电影只有340位演员。由于每个临时演员都与参演该电影的其余1 287个演员相连，从而在$k=1\ 287$处形成一个局部峰值。正是因为这个峰值存在，用幂律分布对度分布进行拟合时无法通过柯尔莫哥洛夫 - 斯米尔诺夫标准。实际上，如表**3-3** 所示，无论是纯幂律分布还是具有高度截断的幂律分布，在拟合演员网络的度分布时都不是统计显著的。然而，这一点并不会改变度分布的幂律本质。

表**3-3** 指数拟合

	λ	k_{min}	p 值	百分比
电网	0.517	4	0.91	12%

对于电网而言，幂律度分布拟合不具有统计显著性。实际上，有很多证据表明，电网

不是无标度的。我们基于本节所描述的拟合方法，用指数函数$e^{-\lambda k}$对电网的度分布进行拟合，得到了具有统计显著性的拟合结果。表中列出了得到的参数λ、最佳拟合时的k_{min}、p值，以及该拟合适用的数据点所占的百分比。

（3） 作为前述问题的结果，用来拟合幂律分布的方法通常会局限在一个很小的标度区间内，使我们需要舍弃很大比例的节点（通常多达99%，参见表**3-4**）来获得统计显著的拟合。一旦将得到的拟合模型绘制在原始数据的旁边，我们会发现拟合结果可能是非常荒谬的，尽管拟合具有统计显著性。

表3-4 真实网络的拟合参数

	$K^{-\gamma}$;（K_{min}, ∞）				$(k+k_{sat})^{-\gamma}e^{-k/k_{cut}}$			
	γ	K_{min}	p值	百分比	γ	k_{sat}	k_{cut}	p值
网络	3.42	72	0.13	0.6%	3.55	8	8 500	0.00
万维网（入度）	2.00	1	0.00	100%	1.97	0	660	0.00
万维网（出度）	2.31	7	0.00	15%	2.82	8	8 500	0.00
电网	4.00	5	0.00	12%	8.56	19	14	0.00
手机通话网络（入度）	4.69	9	0.34	2.6%	6.95	15	10	0.00
手机通话网络（出度）	5.01	11	0.77	1.7%	7.23	15	10	0.00
电子邮件网络（入度）	3.43	88	0.11	0.2%	2.27	0	8 500	0.00
电子邮件网络（出度）	2.03	3	0.00	1.2%	2.55	0	8 500	0.00
科学合作网络	3.35	25	0.000 1	5.4%	1.50	17	12	0.00
演员网络	2.12	54	0.00	33%	—	—	—	0.00
引文网络（入度）	2.79	51	0.00	3.0%	3.03	12	5 691	0.69
引文网络（出度）	4.00	19	0.00	14%	−0.16	5	10	0.00
代谢网络（入度）	2.43	3	0.00	57%	3.85	19	12	0.00
代谢网络（出度）	2.90	5	0.00	34%	2.56	15	10	0.00
酵母蛋白质相互作用网络	2.89	7	0.67	8.3%	2.95	2	90	0.52

表中给出了本书中使用的各个参考网络的估计度指数及最佳拟合参数。我们使用了两种拟合策略。第一种是在(K_{min}, ∞)区域上用纯幂律分布拟合，第二种是用具有低度饱和和指数截断的幂律分布对整个数据集进行拟合。表中列出了具有最佳统计显著性的拟合所对应的度指数γ和K_{min}、最佳拟合的p值，以及最佳拟合适用的数据点所占的百分比。在第二种情况下，我们列出了度指数γ、两个拟合参数k_{sat}和k_{cut}，以及拟合的p值。注意，这里认为$p>0.01$是统计显著的。

综上所述，估计度指数仍然是一个尚未成熟的技术。我们依然缺乏既可以得到统计显著性，又可以在实践中被接受的方法。盲目地应用前面所介绍的工具，要么会导致显然不符合数据趋势的拟合，要么会得到对幂律分布假设的错误拒绝。一个重要的进步是，我们提升了

推导度分布期望形式的能力。我们将在第5章 讨论这个问题。

3.14 参考文献

[1] H. Jeong, R. Albert, and A.-L. Barabási. Internet: Diameter of the world-wide web. *Nature* , 401:130-131, 1999.

[2] A.-L. Barabási and R.Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[3] V. Pareto. *Cours d'Économie Politique: Nouvelle édition* par G.- H. Bousquet et G. Busino, Librairie Droz, Geneva, 299-345, 1964.

[4] A.-L. Barabási. *Linked: The New Science of Networks.* Plume, New York, 2002.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On powerlaw relationships of the internet topology. Proceedings of SIGCOMM. *Comput. Commun* . *Rev* . 29: 251-262, 1999.

[6] R. Pastor-Satorras and A.Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach.* Cambridge University Press, Cambridge, 2004.

[7] D. J. De Solla Price. Networks of Scientific Papers. *Science* , 149: 510- 515, 1965.

[8] S. Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *Eur. Phys. J. B* , 4: 131, 1998.

[9] R. Kumar, P. Raghavan, S. Rajalopagan, and A.Tomkins. Extracting Large-Scale Knowledge Bases from the Web. Proceedings of the 25th VLD BConference, Edinburgh, Scotland, pp.639-650, 1999.

[10] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory of scale- free random networks. *Physica A* , 272:173-187, 1999.

[11] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L.Barabási. The large-scale organization of metabolic networks. *Nature* , 407: 651-654, 2000.

[12] A. Wagner, A. and D.A. Fell. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* , 268: 1803-1810, 2001.

[13] W. Aiello, F. Chung, and L.A. Lu. Random graph model for massive graphs, Proc. 32nd ACM Symp. Theor. Comp, 2000.

[14] H. Jeong, B. Tombor, S. P. Mason, A.-L. Barabási, and Z.N.Oltvai. Lethality and centrality in protein networks. *Nature* , 411:41-42, 2001.

[15] A. Wagner. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. B* , 270: 457-466, 2003.

[16] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci* . 98: 404-409, 2001.

[17] A.-L. Barabási, H. Jeong, E. Ravasz, Z. Néda, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A* , 311: 590-614, 2002.

[18] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Aberg. The Web of Human Sexual Contacts. *Nature* , 411: 907-908, 2001.

[19] R. Ferrer i Cancho and R.V. Solé. The small world of human language. *Proc. R. Soc. Lond. B* , 268: 2261-2265, 2001.

[20] R. Ferrer i Cancho, C. Janssen, and R.V. Solé. Topology of technology graphs: Small world patterns in electronic circuits. *Phys. Rev. E* , 64: 046119, 2001.

[21] S. Valverde and R.V. Solé. Hierarchical Small Worlds in

Software Architecture. arXiv:cond-mat/0307278, 2003.

[22] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E* , 66: 035103(R), 2002.

[23] J.P.K. Doye. Network Topology of a Potential Energy Landscape: A Static Scale-Free Network. *Phys. Rev. Lett* . 88:238701, 2002.

[24] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabó, D. Lazer, K.Kaski, J. Kertesz, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* , 104: 7332-7336 （ 2007 ） .

[25] H. Kwak, C. Lee, H. Park, S. Moon. What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World Wide Web, 591-600, 2010.

[26] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. Proceedings of international AAAI Conference on Weblogs and Social, 2010.

[27] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The Anatomy of the Facebook Social Graph. ArXiv:1111.4503, 2011.

[28] L.A.N. Amaral, A. Scala, M. Barthelemy and H.E. Stanley. Classes of small-world networks. *Proceeding National Academy of Sciences U. S. A* . 97:11149-11152, 2000.

[29] R. Cohen and S. Havlin. Scale free networks are ultrasmall. *Phys. Rev. Lett* . 90, 058701, 2003.

[30] B. Bollobás and O. Riordan. The Diameter of a Scale-Free Random Graph. *Combinatorica* , 24: 5-34, 2004.

[31] R. Cohen and S. Havlin. *Complex Networks - Structure,Robustness and Function.* Cambridge University Press,Cambridge, 2010.

[32] K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett* . 87:278701, 2001.

[33] F. Karinthy. Láncszemek, in *Minden másképpen van* .Budapest, Atheneum Irodai es Nyomdai R.-T. Kiadása, 85-90, 1929. English translation in: M.E.J. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks.* Princeton University Press, Princeton, 2006.

[34] P.S. Dodds, R. Muhamad and D.J. Watts. An experimental study to search in global social networks. *Science* , 301: 827-829, 2003.

[35] P. Erdös and T. Gallai. Graphs with given degrees of vertices. *Matematikai Lapok* , 11:264-274, 1960.

[36] C.I. Del Genio, H. Kim, Z. Toroczkai, and K.E. Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLOS ONE* , 5: e10012, 04 2010.

[37] V. Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat* ., 80: 477-480, 1955.

[38] S. Hakimi. On the realizability of a set of integers as degrees of the vertices of a graph. *SIAM J.Appl. Math* ., 10: 496-506, 1962.

[39] I. Charo Del Genio, G. Thilo, and K.E. Bassler. All scale-free networks are sparse. *Phys. Rev. Lett* . 107: 178701, 10 2011.

[40] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J.Combin* . 1: 311-316, 1980.

[41] M. Molloy and B. A. Reed. Critical Point for Random Graphs with a Given Degree Sequence. *Random Structures and Algorithms* , 6: 161-180, 1995.

[42] M. Newman. *Networks: An Introduction* . Oxford University, Oxford, 2010.

[43] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science* , 296: 910-913, 2002.

[44] G. Caldarelli, I. A. Capocci, P. De Los Rios, and M.A.Muñoz. Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Phys. Rev. Lett* . 89: 258702, 2002.

[45] B. Söderberg. General formalism for inhomogeneous random graphs. *Phys. Rev. E* , 66: 066121, 2002.

[46] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Phys. Rev. E* , 68: 036112, 2003.

[47] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Powerlaw distributions in empirical data. *SIAM Review* , S1: 661-703, 2009.

[48] S. Redner. Citation statistics from 110 years of physical review. *Physics Today* , 58: 49, 2005.

第4章 巴拉巴西－阿尔伯特模型

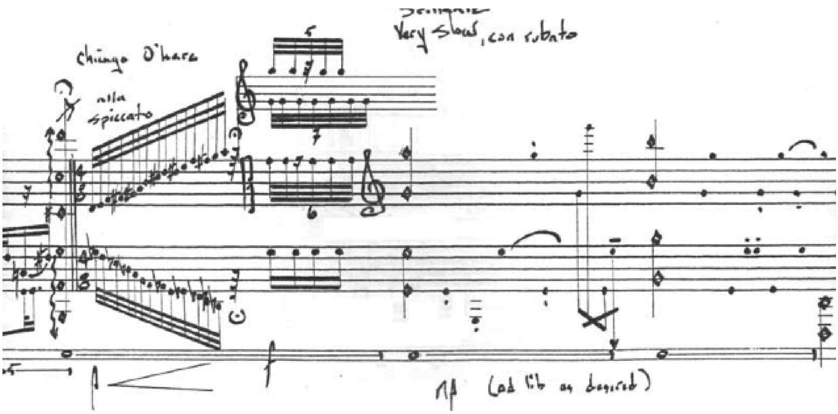


图4-0　无标度奏鸣曲

迈克尔·爱德华·埃杰顿（Michael Edward Edgerton）于2003年创作的《1号钢琴奏鸣曲》（*1 sonata for piano*　）结合生长和偏好连接机制来模拟无标度网络的诞生。图中展示了被埃杰顿称为5号枢纽节点的形成之初。这位作曲家是这样解释音乐和网络之间的关系的："6个不同长度和音程的枢纽节点分布在第二乐章和第三乐章。从音乐角度讲，这6个枢纽节点就像将所有交通流引向一个狭小着陆空间的机场一样。"

本章索引

4.1　本章概览 / 193

4.2　生长和偏好连接 / 195

4.3　巴拉巴西 - 阿尔伯特模型 / 199

4.4　度动力学 / 202

4.5　度分布 / 205

4.6　生长或偏好连接的缺失 / 207

4.7　测量偏好连接 / 211

4.8　非线性偏好连接 / 214

4.9　偏好连接的起源 / 217

4.10　直径与集聚系数 / 225

4.11　小结 / 227

4.12　课后习题 / 229

4.13　进阶阅读4.A
推导度分布 / 231

4.14　进阶阅读4.B
非线性偏好连接 / 235

4.15　进阶阅读4.C
集聚系数 / 239

4.16　参考文献 / 241

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

萨拉·莫里森 **SARAH MORRISON**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

阿迈勒·侯赛尼 **AMAL HUSSEINI**

莫罗·马蒂诺 **MAURO MARTINO**

菲利普·赫费尔 **PHILIPP HOEVEL**

罗伯塔·西纳特拉 **ROBERTA SINATRA**

4.1
本章概览

枢纽节点是无标度网络和随机网络之间最显著的差异。在万维网中，枢纽节点是那些拥有大量指入链接的网站，例如google.com和facebook.com；在代谢网络中，枢纽节点是那些参与大量化学反应的能量载体，例如ATP和ADP。枢纽节点的存在以及与之紧密相关的无标度拓扑，向我们提出了两个基本问题：

（1） 万维网和细胞之间差异如此之大，为什么却具有相似的无标度网络结构呢？

（2） 为什么埃尔德什－雷尼的随机网络模型不能产生真实网络中出现的枢纽节点和幂律分布呢？
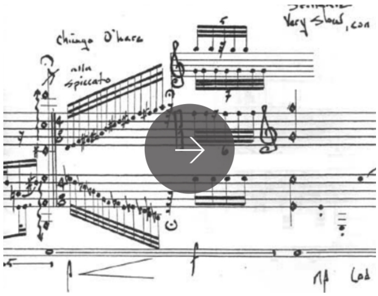
第一个问题特别令人费解，这些性质、起源和范畴完全不同的系统都具有无标度特性：

（1） 细胞网络中的节点是蛋白质或代谢物，而万维网中的节点是储存信息的文档。

（2） 细胞中的链接是化学反应和相互作用，而万维网中的链接是URL（统一资源定位符）或一小段计算机代码。

（3） 二者之间最大的不同是它们的历史：细胞网络是经过了40亿年的进化形成的，万维网却只有几十年的历史。

（4） 代谢网络的作用是为细胞存活提供必要的化学成分，万维网则是为了获取和传递信息而建立的。



在线资源4.1　《1号钢琴奏鸣曲》

埃杰顿的《1号钢琴奏鸣曲》的录音，该曲的创作受到了无标度网络的启发。

扫描二维码下载“湛庐阅读”App，回复书名，欣赏这首美妙的乐曲。

为了弄清楚为什么如此不同的系统会最终形成类似的结构，我们首先需要理解无标度特性的产生机制，这正是本章的主题。由于这些具有无标度特性的系统之间是如此的千差万别，无标度特性的产生一定有一个简单而基本的解释。而这个解释将会改变我们建模网络的方式，并最终指引我们从刻画网络拓扑结构转向建模复杂系统的演化。

4.2
生长和偏好连接

我们对无标度特性的探索之旅始于这样一个问题：为什么随机网络中没有枢纽节点和幂律度分布呢？答案出现在1999年，对于这个问题的回答指出了隐藏在埃尔德什 - 雷尼模型背后的两个假设，而这两个假设违背了真实网络的性质[1]。接下来，我们将分别讨论这两个假设。

网络通过添加新节点而不断生长

随机网络模型假设节点的数目N是固定的。然而，在真实网络中，由于新节点的加入，节点的数目是不断增长的。让我们考虑以下几个例子：

- 1991年，万维网中只有一个节点，就是由万维网创立者蒂姆·伯纳斯 - 李建立的第一个网页。今天，万维网中包含超过1兆（10^{12}）个网页，这个巨大的数字是由数百万个人和机构不断增加新网页而最终达到的（图**4-1a**）。
- 随着新研究论文的发表，科学合作网络和引文网络不断扩大（图**4-1b**）。
- 好莱坞演员网络随着新电影的发行而不断扩大（图**4-1c**）。

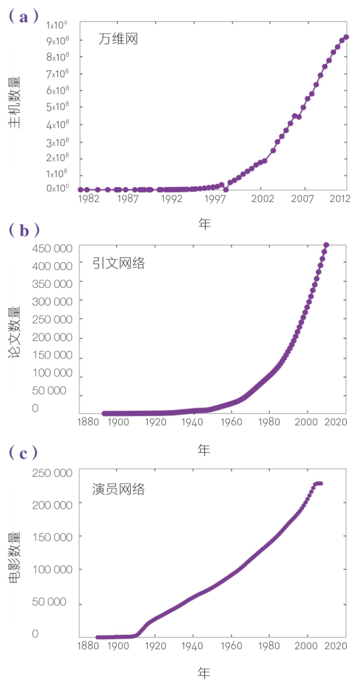


图**4-1**　网络生长

网络不是静止的，而是以增加新节点的方式在增长：

（**a**）万维网主机数量的演变记录了万维网的快速增长。

（**b**）《物理评论》（*Rhysical Review*）自创办以来发表的科研论文数量。如图所

示，论文数量的增长带动了科学合作网络和引文网络的增长。

（c）互联网电影数据库（IMDB.com）中列出的电影数量的增长，同时推动了演员网络的增长。

- 细胞内的蛋白质相互作用网络看上去好像是静态的：我们的基因和蛋白质从父母那里继承而来。然而，这些网络也不是静态的：在过去40多亿年中，基因的数量从几个增长了到了现在人类细胞中的2万多个。

因此，如果想建模这些网络，就不能采用静态模型。相反，我们的建模方法必须承认这样一个事实：网络是一个增长过程的产物。

节点倾向于与链接数高的节点相连

随机网络模型假设节点随机地选择其他节点进行连接。然而，在大多数真实网络中，新加入的节点更倾向于与链接数高的节点相连，这一过程被称为偏好连接（图**4-2**）。

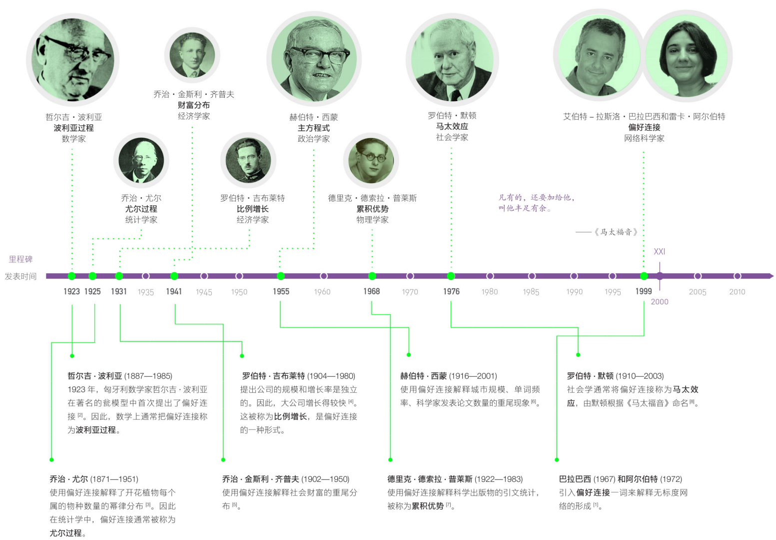


图**4-2**　偏好连接简史

我们来考虑几个例子：

- 在万维网数以万亿计的网页中，我们可能只知道其中很少的一部分。我们所知道的节点并不是完全随机的：我们都知道谷歌和Facebook，却很少会访问充斥在万维网中的数十亿不怎么出名的网页。相比于链接数较低的节点，我们对链接数高的节点了解得更多，因此也更有可能连接到一个链接数高的节点。
- 每年有100多万篇科学论文发表，没有哪个科学家可以读完所有这些论文。一篇论文被引用的次数越多，我们就越有可能听说这篇论文并最终阅读它。由于我们只会引用阅读过的论文，我们的引用便会更倾向于引用数高的论文——引文网络中的大度节点。
- 一个演员所出演的电影越多，导演就越熟悉他的表演技巧。因此，一个演员在演员网络中的度越高，就越有可能获得出演新角色的机会。

总之，随机网络和真实网络之间有两个重要的不同之处：

1.生长

真实网络是一个生长过程的结果，因此其节点数N 会持续增加。相反，随机网络模型假设节点数N 是固定的。

2.偏好连接

真实网络中，新节点倾向于和链接数高的节点相连。相反，随机网络中的节点随机地选择节点进行连接。

真实网络和随机网络还有许多不同之处，后面的章节会详细讨论。然而，正如我们接下来将看到的，生长与偏好连接对于网络的度分布具有非常重要的影响。

很多学科都独立地发现了偏好连接，用以帮助解释不同系统中表现出的幂律特征。在网络科学中，偏好连接于1999年被提出，用于解释无标度性质。

4.3 巴拉巴西 - 阿尔伯特模型

在认识到真实网络中生长和偏好连接共同存在之后，我们提出了一个最简单的、可以生成幂律度分布的网络生长模型。该模型被称为巴拉巴西 - 阿尔伯特模型，也被称为BA模型或者无标度模型。模型的定义如下：

初始时，网络中有m_0个节点，这些节点之间任意连接，只需保证每个节点至少有一个链接即可。接下来，网络按照下面两个步骤演变（图**4-3**）：

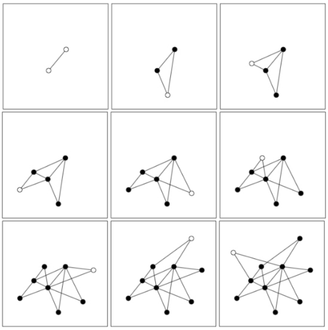


图**4-3**　巴拉巴西 - 阿尔伯特模型的演化

这组图片展示了巴拉巴西 - 阿尔伯特模型连续9个时间步的生长情况。空心圆圈表示网络中新加入的节点，这些节点根据偏好连接机制（公式**4.1**）选择将它的两条链接（$m=2$）连向网络中存在的其他节点。[9]

1.生长

每个时间步，我们向网络中添加一个拥有m（$\leq m_0$)条链接的新节点，该节点会和网络中已经存在的m个节点相连。

2.偏好连接

新节点每次在选择节点进行连接时，选择度为k_i的节点进行连接的概率为：

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \quad (4.1)$$

偏好连接是一个概率化的机制：新节点可以和网络中的任意节点相连，无论对方是枢纽节点还是只有单个链接的节点。但是，如公式**4.1** 所示，在一个度为2的节点和一个度为4的节点之间做选择时，新节点选择与度为4的节点相连的概率是其选择与度为2的节点相连的概率的2倍。

经过t 个时间步后，巴拉巴西 - 阿尔伯特模型生成了一个网络节点数为$N=t+m_0$、链接数为$N=m_0+mt$ 的网络。如图**4-4** 所示，生成的网络具有幂律度分布，度分布的幂指数是$\gamma=3$。关于巴拉巴西 - 阿尔伯特模型的数学分析，参见边栏**4.1** 。

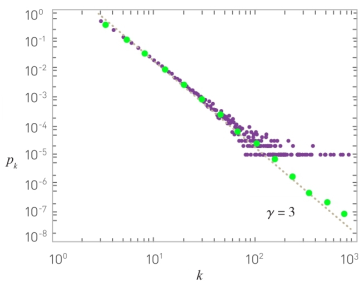


图**4-4** 度分布

由巴拉巴西 - 阿尔伯特模型生成的网络的度分布。图中展示了一个$N=100\ 000$、$m=3$的网络的度分布p_k。紫色表示线性分箱的p_k，绿色表示对数分箱的p_k。图中的直线用于辅助观察，其斜率为$\gamma=3$，表示该网络的度指数。

边栏**4.1**

巴拉巴西 - 阿尔伯特模型的数学定义

巴拉巴西 - 阿尔伯特模型的定义中有许多数学细节没有指明：

- 未指明初始的m_0 个节点之间的连接情况；
- 未指明新节点的m 条链接是逐个加入的还是同时加入的。这有时候会导致潜在的数学冲突：如果链接之间是彼此独

立的，那么这些链接可以连接到同一个节点，从而形成多重链接。

波罗巴斯及其合作者 [10] 提出了线性化弦图（LCD）来解决这些问题，使该模型更适用于数学推导。

根据LCD，我们构造了一个$m=1$时的图$G_1^{(t)}$（图4-5），构造方式为：

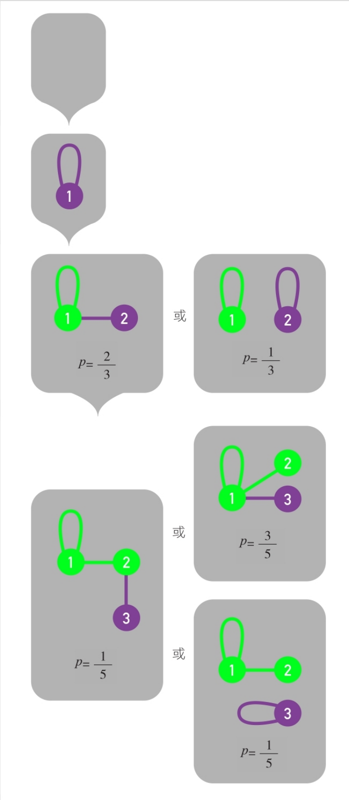


图4-5 线性化弦图（LCD）

线性化弦图为巴拉巴西－阿尔伯特模型提供了一个适合进行精确数学计算的版本 [10]。图中展示了$m=1$时网络演化过程的前4步：

$G_1^{(0)}$：我们从一个空网络开始。

$G_1^{(1)}$：第一个节点只能和自己相连，形成一个自环。自环是允许的。当$m>1$时，多重链接也是允许的。

$G_1^{(2)}$ ：节点2以概率2/3连接到节点1，以概率1/3连接到自己。根据公式**4.2** ，节点2带来的边有一半算作已经存在的边。因此，节点1的度为$k_1=2$，节点2的度为$k_2=1$，归一化常数为3。

$G_1^{(3)}$ ：假设$G_1^{(2)}$ 在实际过程中选择了第一种情形，当节点3加入时，它有三个选择：以概率1/5连接到节点2，以概率3/5连接到节点1，或者以概率1/5连接到自己。

（**1**）从$G_1^{(0)}$ 开始，$G_1^{(0)}$ 是一个没有节点的空图；

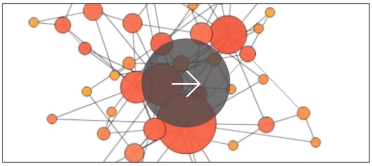
（**2**）通过向$G_1^{(t-1)}$ 中添加节点v_t 以及一条连接v_t 和v_i 的链接而生成图$G_1^{(t)}$ 。节点v_i 按照如下概率被选择：

$$p=\begin{cases}\dfrac{k_i}{2t-1} & \text{当 } 1\leqslant i\leqslant t-1 \\ \dfrac{1}{2t-1} & \text{当 } i=t\end{cases} \text{时} \quad (4.2)$$

也就是说，来自新节点v_t 的链接以概率$k_i/(2t-1)$连接到节点v_i ，此时新链接对节点v_t 的度也做出了贡献。节点v_t 也以概率$1/(2t-1)$连接到自己，即公式**4.2** 中的第二项。注意，该模型允许自环和多重链接出现。不过，当$t\to\infty$时，自环和多重链接的数目可以忽略不计。

当$m>1$时，构造$G_m^{(t)}$ 的过程中新节点v_t 的m 条链接是逐个加入的。加入每条链接时，允许多重链接出现。

如图**4-3** 和在线资源**4.2** 所示，虽然网络中大部分节点只有几个链接，一些节点却逐步成长为枢纽节点。这些枢纽节点就是“富者愈富”现象的结果：偏好连接机制使得新节点倾向于与链接数高的节点相连。因此，链接数高的节点会比链接数低的节点更容易获得新链接，最终发展成为枢纽节点。



在线资源**4.2**　无标度网络的涌现

该视频展示了无标度网络的生长情况和巴拉巴西 - 阿尔伯特模型中枢纽节点的形成过程。视频由王大顺（Dashun Wang）提供。

扫描二维码下载“湛庐阅读”App，回复书名，了解关于无标度网络的更多内容。

总之，巴拉巴西 - 阿尔伯特模型表明，生长和偏好连接这两个简单的机制是无标度网络形成的原因。幂律分布和枢纽节点源于这两个机制引起的“富者愈富”现象。为了理解该模型的行为并刻画无标度性质的涌现，我们需要熟悉该模型的数学性质，这将是下一章的主题。

4.4 度动力学

为了理解无标度特性的涌现，我们来考察巴拉巴西 - 阿尔伯特模型随时间演变的过程。首先，我们考察单个节点的度随时间的变化情况[11]。

在巴拉巴西 - 阿尔伯特模型中，每当有新节点加入时，网络中已经存在的节点就有机会增加它的度。新节点加入网络时，它会从网络中已经存在的$N(t)$个节点中选择m个与之连接。新节点每次选择节点i与之相连的概率由公式**4.1**给出。

我们用一个连续实数变量来近似k_i，该变量可以理解为k_i在很多次网络生长过程中的平均值。那么，节点i获得新链接的速率可以写成：

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j} \qquad (4.3)$$

系数m 体现的是每个新节点会带来m 个链接。因此，节点i 有m次被选择的机会。公式**4.3** 的分母中出现的求和项，是关于网络中除新加入节点之外的所有节点进行的。

因此有

$$\sum_{j=1}^{N-1} k_j = 2mt - m \qquad (4.4)$$

从而，公式**4.4** 可以写为：

$$\frac{dk_i}{dt} = m\Pi(k_i) = \frac{k_i}{2t-1} \qquad (4.5)$$

当t 取值很大时，分子中的-1项可以忽略不计，从而得到：

$$\frac{dk_i}{k_i} = \frac{1}{2}\frac{dt}{t} \qquad (4.6)$$

对公式**4.6** 进行积分，并且考虑到$k_i(t_i) = m$ （节点i 在时刻t_i 加入网络时拥有m 个链接），我们有

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad (4.7)$$

我们称β 为动态指数，取值为

$$\beta = \frac{1}{2}$$

从公式**4.7** 中可以得出如下推论：

（**1**） 每个节点的度都依照幂律增长，并且有相同的动态指数$\beta = 1/2$（如图**4-6a** ）。因此，所有节点遵循相同的增长规律。

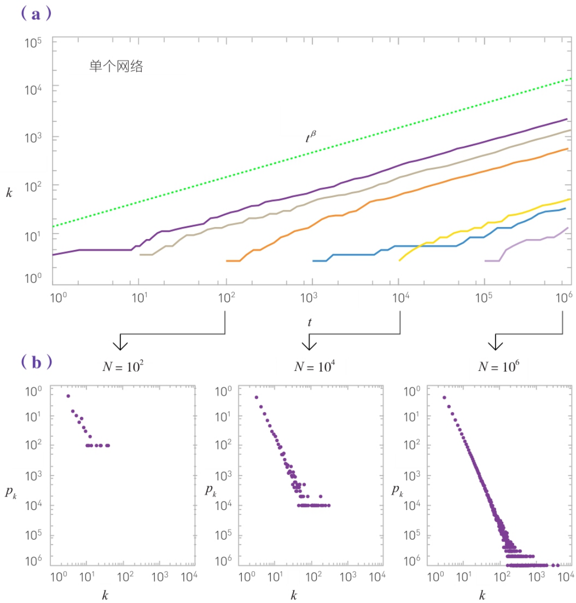


图4-6　度动力学

（a）巴拉巴西 - 阿尔伯特模型中，在时刻$t=1, 10, 10^2$，10^3，10^4，10^5 加入的节点对应的度增长情况（对应自左向右的连续直线）。节点的度按照公式**4.7** 增长。因此，在任意时刻，较早加入的节点会有更大的度。虚线对应公式**4.7** 在$\beta=1/2$时的解析解。

（b）在添加了$N=10^2$，10^4，10^6 个节点后网络的度分布，分别对应时刻$t=10^2$，10^4，10^6 （对应关系由图a中的箭头指示）。网络越大，度分布的幂律特性越明显。注意，为了更好地观察无标度逐渐出现的过程，我们使用了线性分箱来计算p_k。

（2）度的增长是亚线性的（即$\beta<1$）。这是巴拉巴西 - 阿尔伯特模型生长特性的必然结果：每个新节点在加入网络时，都比在它之前加入网络的节点有更多的节点可以连接。因此，随着时间的积累，每个节点会与更多的节点竞争被连接的机会。

（3）节点i 越早加入网络，它的度$k_i(t)$就越高。因此，枢纽节点之所以度很高，是因为它们比其他节点更早加入网络。这种现象在营销和商业领域被称为先发优势。

（4）节点i 获得新链接的速率可以由公式**4.7** 的导数得到：

$$\frac{dk_i(t)}{dt}=\frac{m}{2}\frac{1}{\sqrt{t_i t}} \qquad (4.8)$$

这意味着在每个时间步，节点越老，获得的链接就越多（因为它们有更小的t_i ）。进一步，节点获得链接的速率随着时间下降$t^{-1/2}$ 。随着时间推移，每个节点获得链接的数目会越来越少。

综上所述，巴拉巴西 - 阿尔伯特模型为我们提供了一个网络演变的动态描述，刻画了“真实网络中节点是逐个加入的”这一事实。该模型中，节点之间形成一种对链接的竞争，老节点比新节点在获得链接方面更有优势，并最终成为枢纽节点。

4.5 度分布

巴拉巴西 - 阿尔伯特模型生成的网络有一个显著的特征——具有幂律度分布（如图**4-4** ）。在这一节中，我们将计算度分布p_k 的函数形式，并理解其形成原因。

有很多分析工具可以用来计算巴拉巴西 - 阿尔伯特网络的度分布。最简单工具是我们在前面的章节 [1] ，[11] 中介绍的连续介质理论。使用该工具得到的度分布为（边栏**4.3** ）：

边栏4.3

连续介质理论

为了在连续介质近似下计算巴拉巴西 - 阿尔伯特模型的度分布，我们首先计算度小于k 的节点数量，即$k_i(t)<k$ 。根据公式**4.7** 可得：

$$t_i < t\left(\frac{m}{k}\right)^{1/\beta} \qquad (4.12)$$

该模型中，每个时间步新加入一个节点（边栏4.2 ）。因此，度小于k 的节点数量为：

$$t\left(\frac{m}{k}\right)^{1/\beta} \quad (4.13)$$

总共有$N = m_0 + t$ 个节点，当t 较大时，有$N \approx t$ 。因此，一个随机选择的节点，其度小于或等于k 的概率记为累积度分布

$$P(k) = 1 - \left(\frac{m}{k}\right)^{1/\beta} \quad (4.14)$$

对公式4.14 求导，得到度分布

$$p_k = \frac{\partial P(k)}{\partial k} = \frac{1}{\beta}\frac{m^{1/\beta}}{k^{1/\beta+1}} = 2m^2k^{-3} \quad (4.15)$$

正好是公式4.9 。

$$p(k) \approx 2m^{1/\beta}k^{-\gamma} \quad (4.9)$$

其中

$$\gamma = \frac{1}{\beta} + 1 = 3 \quad (4.10)$$

因此，度分布服从度指数为$\gamma = 3$的幂律分布，这和数值模拟的结果相吻合（图4-4 和图4-7 ）。而且，公式4 .10 将刻画网络拓扑的量——度指数γ 和刻画节点随时间演化的动态指数β 联系在一起，揭示了网络拓扑和网络动力学之间的深层关系。

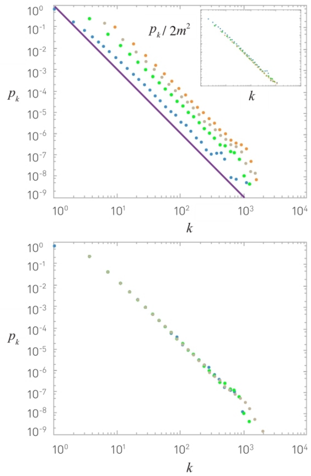


图4-7　解析解探究

（a） 我们生成多个大小为$N=100\ 000$的网络，分别为$m_0=m=1$（蓝色）、3（绿色）、5（灰色）和7（橙色）。可以看到，这些线是相互平行的，表明度指数γ 独立于m_0和m。紫色直线的斜率为-3，对应着解析解预测的度指数$\gamma=3$。子图：公式**4.11** 预测$p_k\sim 2m^2$，因此$p_k/2m^2$应该独立于m。实际上，从图中$p_k/2m^2$关于k的曲线可以看出，这些曲线重叠成了一条曲线。

（b） 在巴拉巴西 - 阿尔伯特模型中，p_k独立于N。为了验证这一点，对于不同的N，我们绘制出相应的p_k。这里，N的取值为$N=50\ 000$（蓝色），100 000（绿色），200 000（灰色），参数$m_0=m=3$。从图中可以看出，不同的N所对应的p_k实际上无法区分，这表明度分布是稳定的，即度分布独立于时间和网络大小。

连续介质理论虽然得到了正确的度指数，但未能准确地预测出公式**4.9** 中幂律度分布的前置因子。正确的前置因子可以通过主方程[12]或速率方程[13]方法得到，也可以使用线性化弦图（LCD）模型精确地计算出来[10]（边栏**4.2**）。如进阶阅读**4.A** 所示，巴拉巴西 - 阿尔伯特模型所得到的度分布，其确切形式为：

边栏4.2

网络中的时间

将真实数据与网络模型的预测进行比较时，我们需要知道如何

度量网络中的时间。真实网络的时间标度差异很大：

万维网

第一个网页创建于1991年。如今，万维网中有数万亿个网页，平均每毫秒（10^{-3} 秒）新增一个节点。

细胞

细胞是40亿年进化的产物。人类细胞中有大约2万个基因，平均下来，细胞网络每20万年（约10^{13} 秒）新增一个节点。

由于网络的时间标度差异巨大，我们不可能使用真实时间来对比不同网络的演化过程。因此，在网络理论中我们使用事件时间——网络拓扑发生一次变化计为一个时间步。

例如，在巴拉巴西 - 阿尔伯特模型中，新增一个节点对应一个时间步，即$t=N$ 。在其他模型中，时间步也可以按照新增链接或删除节点来确定。如果需要的话，我们可以在事件时间和物理时间之间建立一个直接映射关系。

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{4.11}$$

公式**4.11** 意味着：

- 当k 很大时，公式**4.11** 可以约简为$p_k \sim k^{-3}$ ，也就是$\gamma=3$。这和公式**4.9**、公式**4.10** 一致。
- 度指数γ 独立于m ，这和数值模拟的结果是吻合的（图**4-7a**）。
- 在真实网络中观察到的幂律度分布描述了不同年龄或规模的系统。因此，一个合适的模型应该得出一个和时间无关的度分布。实际上，根据公式**4.11**，巴拉巴西 - 阿尔伯特模型得到的度分布是独立于t 和N 的。因此，该模型得到的是一个稳定的无标度状态。数值模拟的结果也证实了这一发现（图**4-7b**），不同的t 和N 所得到的p_k 完全重合。

- 公式**4.11** 推出，幂律分布的系数正比于$m(m+1)$（对于大的m，可以记为m^2）。这一结论同样被数值模拟的结果证实（如图**4-7** 的子图）。

总之，解析计算的结果表明，巴拉巴西 - 阿尔伯特模型生成了度指数为$\gamma=3$的无标度网络。度指数和模型参数m、m_0都无关。而且，模型产生的度分布是稳定的（不随时间而变化），这解释了为什么不同历史、规模和年龄的网络具有相似的度分布。

4.6 生长或偏好连接的缺失

巴拉巴西 - 阿尔伯特模型中生长和偏好连接的并存，引出了一个重要的问题：这两个机制对于无标度特性的出现都是必要的吗？换言之，只保留其中一个机制是否也能生成一个无标度网络呢？为了回答这些问题，下面我们讨论这个模型的两个极限情况，每个情况只包含其中一种机制[1]，[11]。

模型A

为了测试偏好连接的作用，我们仅保留网络的生长机制（要素A），而去掉偏好连接机制（要素B）。具体而言，模型A从m_0个节点开始，按照如下步骤演化：

1.生长

在每个时间步，我们添加一个拥有m条链接的新节点，使之与网络中已经存在的m个节点相连。

2.随机连接

一个新节点和度为k_i的节点相连的概率为：

$$\Pi\left(k_i\right)=\frac{1}{\left(m_0+t-1\right)} \quad (4.16)$$

也就是说，$\Pi(k_i)$和节点度k_i无关。这表明，新节点是随机选择节点进行连接的。

根据连续介质理论可以推出，模型A中$k_i(t)$随时间呈对数增长：

$$k_i\left(t\right)=m\ln\left(e\frac{m_0+t-1}{m_0+t_i-1}\right) \quad (4.17)$$

这比公式**4.7**中的幂律增长要慢得多。模型所得的度分布服从指数分布（图**4-8a**）：

$$p\left(k\right)=\frac{e}{m}\exp\left(-\frac{k}{m}\right) \quad (4.18)$$

指数函数要比幂函数的衰减快得多，此时网络中很难出现枢纽节点。因此，缺少偏好连接的模型无法得到无标度特性和枢纽节点。实际上，当所有节点获得链接的概率相同时，我们缺少了“富者愈富”的过程，因此没有明显的“胜者”。

模型B

为了测试生长机制的作用，接下来我们在网络中保留偏好连接（要素B）而去掉生长机制（要素A)。模型B从N个节点开始，按照如下步骤演化：

1.偏好连接

在每个时间步，随机选择一个节点，该节点和其他节点相连时，选择度为k_i的节点的概率可根据公式**4.1**得到。由于$\Pi(0)=0$，我们假设度$k=0$的节点的度为$k=1$，否则这些节点无法获得链接。

在模型B中，网络中的节点数目在网络演化过程中保持不变，而链接数随时间呈线性增长。因此，当t很大时，每个节点的度也随时间呈线性增长（图**4-7b**的子图）：

$$k_i(t) \approx \frac{2}{N} t \quad (4.19)$$

实际上，在每个时间步，我们添加一个新的链接，而不改变节点的数量。

在早期，网络中只有少数几个链接（即$L \ll N$），此时每个新链接会连接两个之前没有任何链接的节点。在这个阶段，模型B的演化和$m=1$时的巴拉巴西 - 阿尔伯特模型没有区别。数值模拟表明，在这个阶段，模型B得到的度分布服从幂律分布（图**4-8b**）。

然而，p_k 不是稳定的。实际上，短暂的一段时间过后，节点度开始收敛到平均度公式**4.19**，度分布中开始出现峰值（图**4-8b**）。当$t \to N(N-1)/2$时，网络变成了一个完全图——图中所有节点的度都是$k_{max}=N-1$，此时有$p_k=\delta(N-1)$。

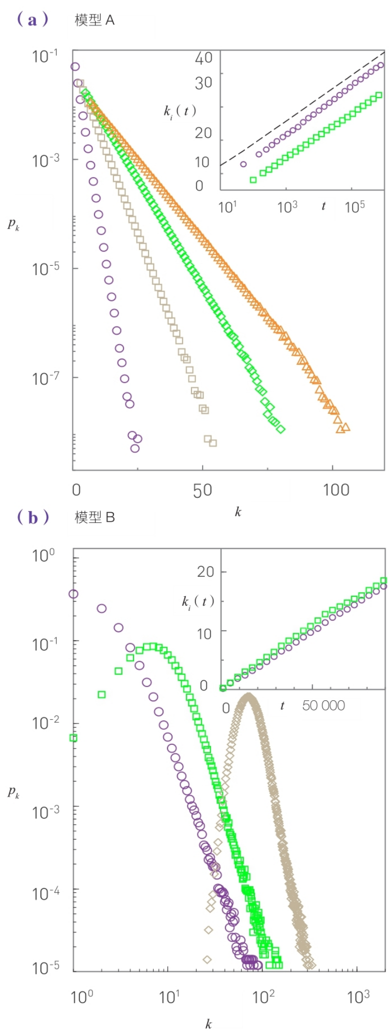


图4-8　模型A和模型B

检验生长机制和偏好连接机制作用的数值模拟。

（**a**）模型**A**

模型A的度分布，该模型包含生长机制，但没有偏好连接机制。图中的不同标记分别对应$m_0=m=1$（圆圈）、3（方块）、5（菱形）和7（三角形），网络大小为$N=800\ 000$。在线性 - 对数坐标系下可以清晰地看出，该模型生成的网络具有指数度分布p_k，这和公式**4.18** 所预测的一致。

子图：两个节点的度演化情况。这两个节点分别是在$t_1=7$和$t_2=97$时加入网络

的，参数$m_0 = m = 3$。虚线对应公式**4.17**。

（**b**）模型**B**

模型B的度分布，该模型包含了偏好连接机制，但没有生长机制。图中所示的是$N = 10\ 000$，$t = N$（圆圈）、$t = 5N$（方块）、$t = 40N$（菱形）时的情形。度分布p_k的形状在一直变化，这表明度分布是不稳定的。

子图：两个节点的度与时间的关系（$N = 10\ 000$）。从中可以看出，$k_i(t)$呈线性增长，和公式**4.19**一致。[11]

综上所述，缺少偏好连接机制的模型A会得到一个稳定的指数度分布。相反，缺少生长机制的模型B无法得到稳定的分布，最终网络收敛到一个完全图。模型A和模型B都无法再现真实网络中观测到的无标度度分布。这表明，生长机制和偏好连接机制对于无标度特性的出现而言都是不可或缺的。

4.7 测量偏好连接

通过前一节我们了解到，生长和偏好连接共同导致了网络的无标度特性。对于真实系统而言，生长是显而易见的：所有大型网络都是通过不断增加新节点而达到如此规模的。然而，要相信真实网络中也存在偏好连接，我们还需要进行实验验证。在本节中，我们将介绍如何通过测量真实网络的$\Pi(k)$函数来检测偏好连接是否存在。

偏好连接依赖于两个假设：

假设**1**

一个节点被连接的概率$\Pi(k)$取决于它的度k。这与随机网络模型截然不同，随机网络模型中的$\Pi(k)$与k无关。

假设**2**

$\Pi(k)$和k是线性关系。

这两个假设可以通过测量$\Pi(k)$来验证。具体而言，在下面两种情况下我们可以测量$\Pi(k)$。第一种情况是，我们知道每个节点

加入网络中的时间。第二种情况是，我们知道一个网络在相隔不远的两个时刻的结构[14]，[15]。

考虑第二种情况，我们知道同一个网络在两个不同时刻的结构。第一个时刻记为t，第一个时刻记为$t+\Delta t$（图**4-9a**）。假设节点i的度在这两个时间间隔内发生了变化，度的变化量记为$\Delta k_i = k_i(t+\Delta t)-k_i(t)$。根据公式**4.1**，相对变化量$\Delta k_i/\Delta t$为：

$$\frac{\Delta k_i}{\Delta t} \sim \Pi(k_i) \quad (4.20)$$

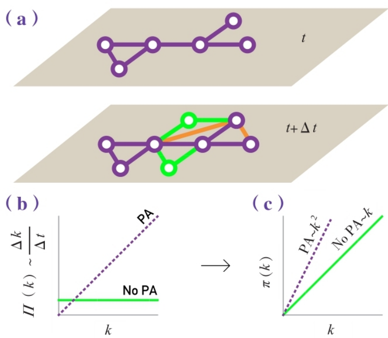


图**4-9**　检测偏好连接

（**a**）如果我们能观测到同一个网络在t和$t+\Delta t$两个时刻的结构，就可以通过将两者进行对比来测量函数$\Pi(k)$。如图所示，两个绿色节点在$t+\Delta t$时加入网络，使网络中部分节点的度发生了变化。橙色的线表示此前不相连的节点之间新增的链接，被称为内部链接，这些链接的作用将在第**5**章中讨论。

（**b**）当偏好连接存在时，$\Delta k/\Delta t$线性依赖于t时刻节点的度。

（**c**）累积偏好连接函数$\pi(k)$的幂指数可以帮助我们检测偏好连接是否存在（图**4-10**）。

这便是偏好连接的函数形式。要使公式**4.20**有效，我们需要保证Δt较小，以便度的变化量Δk不大。然而，Δt也不能过于小，至少要保证这两个时刻的网络之间有可以观察到的变化。实际上，得到的$\Delta k_i/\Delta t$曲线可能包含噪声。为了减少噪声的影响，我们改为测量累积偏好连接函数：

$$\pi(k)=\sum_{k_i=0}^{k}\Pi(k_i) \tag{4.21}$$

当偏好连接不存在时，$\Pi(k_i)$是常数。因此，根据式公式**4.21**，$\pi(k)\sim k$。如果存在线性偏好连接，即$\Pi(k_i)=k_i$，我们有$\pi(k)\sim k^2$。

图**4-10** 展示了4个真实网络中测量到的$\pi(k)$。对于每个网络，我们都可以观察到，$\pi(k)$的增长比线性增长要快，这意味着偏好连接是存在的。图**4-10** 还表明，$\Pi(k)$可以近似为：

$$\Pi(k)\sim k^{\alpha} \tag{4.22}$$

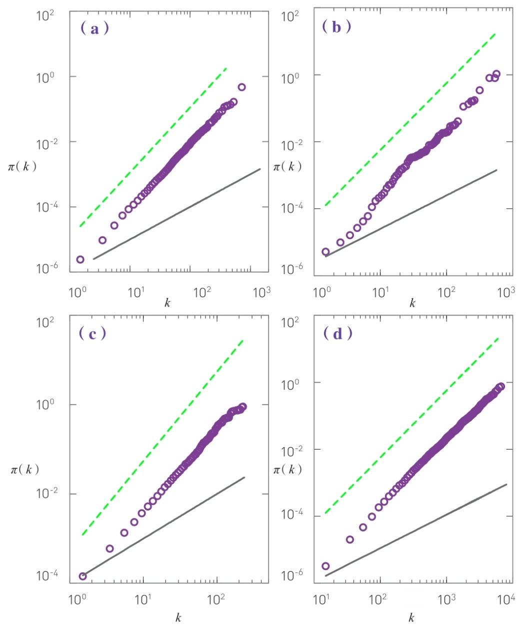


图**4-10**　偏好连接的存在证据

图中展示了几个真实系统中由公式**4.21** 定义的累积偏好连接函数$\pi(k)$：

（**a**）引文网络。

（**b**）互联网。

（**c**）神经科学领域的科学合作网络。

（**d**）演员网络。

每个子图中有两条辅助线：虚线对应线性偏好连接的情况（$\pi(k)\sim k^2$），实线表示不存在偏好连接的情况（$\pi(k)\sim k$）。与假设1一致，我们在每个数据集中都发现了$\Pi(k)$与k的依赖关系。不过，在图c和图d中，$\pi(k)$生长得比k^2慢，表明这些系统中的偏好连接是亚线性的，这与假设2矛盾。注意，这些测量只考虑了新节点加入带来的链接，没有考虑新增的内部链接。[14]

对于互联网和引文网络而言，$\alpha\approx1$，即$\Pi(k)$和k是线性关系，符合公式**4.1**。这与假设1与假设2都是一致的。对于科学合作网络和演员网络而言，最佳拟合结果是$\alpha=0.9\pm0.1$，这意味着出现了亚线性的偏好连接。

总之，公式**4.20**可以帮助我们检测真实网络中是否存在偏好连接。测量的结果表明，偏好连接概率依赖于节点的度。我们还发现，在有些系统中，偏好连接是线性的；而在另一些系统里，偏好连接是亚线性的。对于非线性偏好连接，我们下一节再讨论。

4.8 非线性偏好连接

在图**4-10**中观察到的亚线性偏好连接引出了一个重要问题：这种非线性对网络结构有什么影响？为了回答这个问题，我们将公式**4.1**中的线性偏好连接替换成公式**4.22**的非线性偏好连接，重新计算此时巴拉巴西－阿尔伯特模型得到的度分布。

当$\alpha=0$时，模型的表现是清楚的：由于缺少偏好连接，该模型退化成了第**4.5**节中讨论的模型A。因此，度分布服从公式**4.17**所示的指数形式。

当$\alpha=1$时，该模型等价于标准的巴拉巴西－阿尔伯特模型，从而得到一个度分布服从公式**4.14**的无标度网络。

接下来，我们考察$\alpha\neq0$和$\alpha\neq1$时的情形。对于任意的α，通过计算度分布p_k，可以发现几种不同的模式[13]（进阶阅读**4.B**）：

亚线性偏好连接（$0<\alpha<1$）

对于任意$\alpha>0$的情形，新节点更倾向于和有较多链接的节点相连。不过，当$\alpha<1$时，这种倾向比较弱，还不足以产生无标度的度分布。这个阶段的度分布服从广延指数分布（**3.10**节）：

$$p_k \sim k^{-\alpha}\exp\left(\frac{-2\mu(\alpha)}{\langle k\rangle(1-\alpha)}k^{1-\alpha}\right) \quad (4.23)$$

其中$\mu(\alpha)$仅仅弱依赖于α。公式**4.23**中的指数截断意味着，亚线性偏好连接限制了枢纽节点的大小和数量。

同时，亚线性偏好连接还会影响网络的最大度k_{max}。根据公式**3.18**，无标度网络的最大度k_{max}是时间的多项式函数。对于亚线性偏好连接，我们有：

$$k_{max} \sim (\ln t)^{1/(1-\alpha)} \quad (4.24)$$

这个对数依赖关系说明，网络最大度的生长速度要远低于多项式增长。这就是$\alpha<1$时枢纽节点要小一些的原因（图**4-11**）。

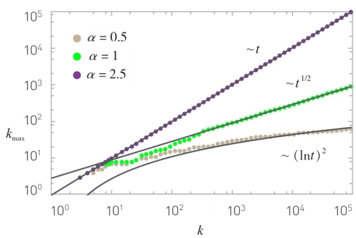


图**4-11**　枢纽节点的增长

偏好连接的性质影响最大节点的度。在无标度网络（$\alpha=1$）中，最大枢纽节点的增速为$t^{1/2}$（绿色线），对应公式**3.18**。对于亚线性偏好连接（$\alpha<1$），最大度的增速是对数形式，对应公式**4.24**。对于超线性偏好连接（$\alpha>1$），最大枢纽节点的度随时间呈线性增长，始终能从新增的链接中分到一定的份额，对应公式**4.25**。图中的标记是数值模拟的结果，虚线表示解析解。

超线性偏好连接（α>1）

当$\alpha>1$时，新节点倾向于和大度节点相连的趋势增强了，加速了“富者愈富”的进程。当$\alpha>2$时，“富者愈富”的结果尤为显著，此时

模型会出现“胜者通吃”的现象：几乎所有节点都连接到一个或几个超级枢纽节点。此时，我们得到了一个中心辐射网络——大部分节点直接连接到几个中心节点。当$1<\alpha<2$时，情况是类似的，只不过没有那么极端罢了。

“胜者通吃”也会影响最大枢纽节点的大小（图**4-11**）：

$$k_{\max} \sim t \qquad (4.25)$$

因此，当$\alpha>1$时，最大枢纽节点和系统中有限比例的节点相连。

总之，非线性偏好连接会改变度分布，要么会限制枢纽节点的大小（$\alpha<1$），要么会导致超级枢纽的产生（$\alpha>1$，图**4-12**）。因此，只有当$\Pi(k)$严格线性依赖于节点的度k时，所得到的网络才具有纯粹的幂律度分布。虽然在很多真实网络中，我们的确观察到了这样的线性依赖关系，但是在诸如科学合作网络和演员网络这些网络中，偏好连接是亚线性的。非线性$\Pi(k)$是真实网络的度分布偏离纯粹幂律分布的一个原因。因此，对于具有亚线性$\Pi(k)$的系统，公式**4.23**中的广延指数分布能够更好地拟合度分布。

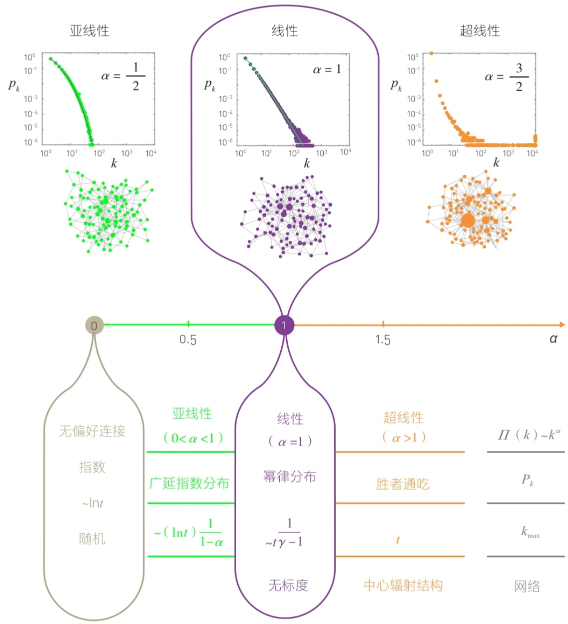


图4-12　非线性偏好连接

非线性巴拉巴西－阿尔伯特模型。顶部三个子图展示的是不同α 下的度分布p_k （$N=10^4$ ），下面对应地展示了$N=100$时的拓扑结构。理论分析表明，会出现4种不同的情况：

无偏好连接（ $\alpha=0$ ）

此时网络具有简单的指数度分布，服从公式**4.18** 。网络中没有枢纽节点，类似于随机网络。

亚线性偏好连接（ $0<\alpha<1$ ）

度分布服从广延指数分布，对应公式**4.23** 。此时，网络中的枢纽节点比无标度网络要少且小一些。当$\alpha\rightarrow 1$时，指数截断的作用消失，度分布p_k 在很大的范围内服从幂律分布。

线性偏好连接（ $\alpha=1$ ）

对应标准的巴拉巴西－阿尔伯特模型，度分布服从幂律分布。

超线性偏好连接（ $\alpha>1$ ）

大度节点的吸引力非常大。“胜者通吃”使网络形成中心辐射结构。在这种情况下，最早加入的几个节点成为超级枢纽节点，后续的节点都连接到它们。图中展示了$\alpha=1.5$时的网络度分布。从中可以看出，大量小度节点和少数几个超级枢纽节点（$k=10^4$ ）并存。

4.9
偏好连接的起源

鉴于偏好连接在真实网络演化中的关键作用，我们一定会问，偏好连接从哪里来？这个问题可以分为两个更具体的问题：

为什么 $\Pi(k)$会依赖于 k ？

为什么 $\Pi(k)$和 k 的依赖关系是线性的？

在过去十年中，关于这些问题，我们见证了思路完全不同的两类回答。第一类认为偏好连接是随机事件和网络结构性质之间相互作用的产物。这些机制不需要网络全局信息，仅依赖于随机事件，因此我们称之为局部或随机机制。第二类回答假设每个新节点或新链接的形成都是在平衡一些相互矛盾的需求，往往取决于成本利益分析。这类模型通常需要知道网络全局信息，且依赖于优化方法，因此我们称之为全局或优化机制。本节我们将讨论这两种不同的机制。

局部机制

巴拉巴西 - 阿尔伯特模型假设存在偏好连接。然而，接下来我们会看到，可以在没有偏好连接的情况下构造出能够生成无标度网络的模型。这类模型可以生成偏好连接。接下来，我们讨论两个这样的模型，并且推导出它们的$\Pi(k)$，以便理解偏好连接的起源。

链接选择模型

链接选择模型或许是能够在没有偏好连接的情况下生成无标度网络的一种最简单的局部机制[16]。它的定义如下（图**4-13**）：

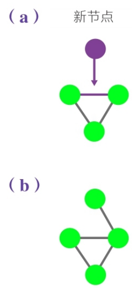


图4-13　链接选择模型

（a）　网络通过添加新节点而生长，每个新节点在加入时会随机选择一条链接（紫色）。

（b）　新节点以相同的概率从该链接的两个端点中选择一个与之相连。在这个例子中，新节点选择了链接的右侧端点。

（1）生长

在每个时间步，我们向网络中添加一个新节点。

（2）链接选择

随机选择一条链接，将新节点连接到所选链接的其中一个端点上。

该模型不需要全局网络拓扑信息，因此本质上属于局部和随机机制。与巴拉巴西-阿尔伯特模型不同，该模型缺少一个内置的$\Pi(k)$函数。然而，接下来我们就会看到，该模型能够产生偏好连接。

首先，一条随机选择的链接，其端点的度为k的概率为：

$$q_k = Ckp_k \tag{4.26}$$

公式4.26刻画了两个效应：

第一，节点的度越高，该节点就越容易成为随机选择的链接的一端。

第二，网络中度为k的节点越多（也就是p_k越高），这些节点越容易成为随机选择的链接的一端。

在公式4.26 中，常数C 可以通过归一化条件$\Sigma q_k=1$计算得到，即$C=1/\langle k\rangle$。因此，一条随机选择的链接其端点的度为k 的概率为：

$$q_k=\frac{kp_k}{\langle k\rangle} \tag{4.27}$$

公式4.27 是新节点和度为k 的节点相连的概率。公式4.27 中，概率q_k 和k 是线性关系。这表明，链接选择模型通过产生偏好连接构造出了无标度网络。

复制模型

链路选择模型提供了最简单的偏好连接机制。在依赖局部机制的模型中，链接选择模型既不是第一个，也不是最流行的一个。这个位置是属于复制模型的（图4-14）。复制模型模拟了一个简单的现象：一个网页的作者倾向于从具有相关内容的其他网页中借鉴链接[17]，[18]。该模型定义如下：

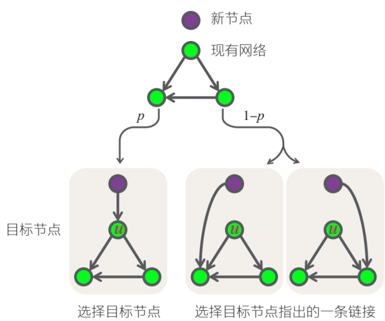


图4-14 复制模型

复制模型的主要步骤。新节点以概率p 连接到一个随机选择的目标节点u ，或者以概率$1-p$ 连接到节点u 指向的某个节点。换言之，新节点以概率$1-p$ 复制了目标节点u 的一条链接。

在每个时间步，一个新节点加入网络中。为了确定新节点会连接到哪个已有节点，我们随机选择一个节点u ；以万维网为例，这类似于选择一个内容和新节点相近的节点。然后，我们进行如下两步过程

（图4-14 ）：

（1） 随机连接：新节点以概率p 连接到节点u ，即我们连接到随机选择的网页。

（2） 复制：我们以概率1-p 随机选择节点u 的一条出边，让新节点连接到这条边的另一端。换言之，新网页复制了节点u 的一条链接并连接到其指向的节点，而不是直接连接到节点u 。

步骤（1）中选择到某个特定节点的概率是$1/N$ 。步骤（2）等价于选择了一个随机选择的链接所指向的节点进行连接。对于无向网络，通过复制步骤（2），选择到度为k 的节点的概率为$k/2L$ 。结合步骤（1）和步骤（2），新节点连接到度为k 的节点的可能性为：

$$\Pi(k)=\frac{p}{N}+\frac{1-p}{2L}k$$

可以看出，$\Pi(k)$和k 是线性关系，对应着线性偏好连接。复制模型的普及得益于其与真实系统的关系：

（1） 社会网络：一个人的朋友越多，他被他的朋友介绍给其他人的机会就越大。换言之，我们"复制"了朋友的朋友。因此，在没有朋友的情况下结交新朋友很困难。

（2） 引文网络：没有哪个科学家能熟知一个领域的所有论文。作者们通过"复制"他们所读论文的参考文献来决定阅读和引用哪些论文。因此，引用次数较多的论文更容易被再次阅读和引用。

（3） 蛋白质相互作用网络：导致细胞中形成新基因的基因复制机制可以映射到复制模型上，以解释蛋白质相互作用网络的无标度性质[19]，[20]。

综上所述，我们发现，链接选择模型和复制模型都可以通过随机连接产生线性偏好连接。

优化模型

经济学一直假设人类做决定的行为是理性的，决定是在权衡了成本与利益之后做出的。换言之，每个人都在最大化个人利益。这是经济学中理性选择理论的基石[21]，也是现代政治学、社会学和哲学的核心假设。接下来我们会看到，这样的理性决定可以带来偏好连接[22]，[23]，[24]。

以互联网为例，其节点是通过电缆连接在一起的路由器。建立一条新的互联网链接需要在两台路由器之间铺设电缆。由于铺设电缆的成本很高，在铺设之前需要仔细进行成本效益分析。每个新加入互联网中的路由器在选择建立链接时，必将在获得良好的网络性能（例如带宽）和电缆建设费用（例如物理距离）之间做出权衡。二者可能是相互冲突的两种诉求，因为最近的节点不一定能够提供最好的网络性能。

简单起见，我们假设所有节点都位于一个单位正方形中。在每个时间步，我们添加一个节点，节点的物理位置通过在正方形内随机选择一个点来确定。新节点i选择节点进行连接时，其成本函数[22]为：

$$C_i = \min_j \left[\delta d_{ij} + h_j\right] \qquad (4.28)$$

成本函数需要考察网络中已经存在的每个节点j。这里，d_{ij}是节点i与节点j之间的欧氏距离，h_j是节点j和“网络中心”之间的“网络距离”。我们把第一个加入网络中的节点视为“网络中心”（图**4-15**），认为其具有最好的网络性能。因此，h_j刻画了节点j能够提供的“资源”，可以用它和“网络中心”之间的“网络距离”来度量。

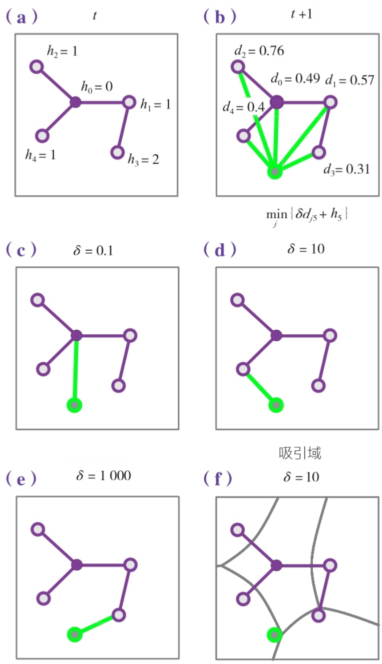


图4-15　优化模型

（a）一个小网络，对于每个节点j，图中给出了公式4.28 中的h_j 项。这里，h_j 表示节点j 到节点$i=0$的“网络距离”，节点$i=0$被指定为“网络中心”，中心节点被认为具有最好的网络性能。因此，$h_0=0$，$h_3=2$。

（b）新节点（绿色）根据公式4.28 选择C_j 最小的节点j 并与之相连。

（c）～（e）当δ 较小时，新节点将连接到$h_j=0$的中心节点。随着δ 增大，公式4.28　中的平衡发生了变化，新节点开始倾向于和距离较远的节点相连。图c～e展示了绿色新节点在不同的δ 取值下做出的选择。

（f）$\delta=10$时每个节点的吸引域。新节点到达时，总是和其所处吸引域的中心节点相连。吸引域的大小取决于中心节点的度。实际上，h_j 越小，越可能吸引到距离它较远的新节点。节点j 的度越大，其与中心节点的期望距离h_j 就越小。

计算表明，根据公式4.28 中δ 的取值及N 的取值，会出现三种不同的网络结构（图4-15 ）：

星型网络$\delta<(1/2)^{1/2}$

当$\delta = 0$时，欧氏距离不起作用。此时，每个节点都和网络中心节点直接相连，从而形成星型网络。实际上，只要h_j 相对于δd_{ij} 而言起主导作用，我们就会得到一个星型网络。

随机网络$\delta \geq N^{1/2}$

当δ 较大时，公式**4.28** 中δd_{ij} 的贡献远大于h_j 。此时，每个新加入的节点都会选择与其欧氏距离最近的节点相连。于是，最终形成的网络的度分布是有界的，类似于随机网络（图**4-16b** ）。

无标度网络$4 \leq \delta \leq N^{1/2}$

数值模拟和解析计算的结果表明，当δ 的取值介于上述两种情况之间时，网络具有无标度拓扑结构[22] 。

在优化模型中，幂律起源于两种相互竞争的机制：

（**1**） 优化：每个节点都有一个吸引域，落在该节点的吸引域中的节点就会和它相连。每个吸引域的大小和处于域中心的节点j 的h_j 有关（图**4-15f** ）。

（**2**） 随机：我们随机选择新节点的位置，落在N 个吸引域之中的某一个。已有节点中，度最大的节点具有最大的吸引域，因此会吸引最多的新节点，得到最多的新链接，最终形成偏好连接（图**4-16d** ）。

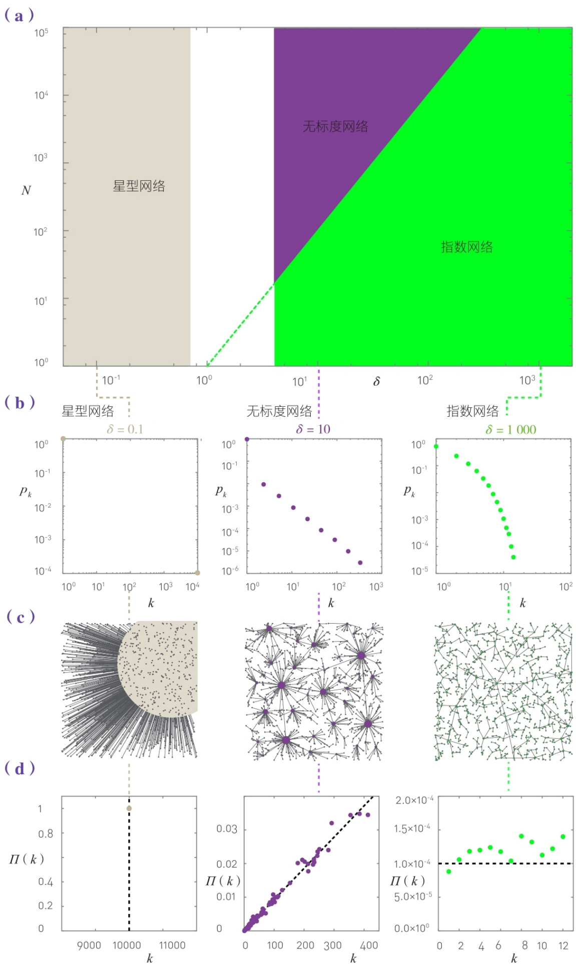


图4-16　优化模型中的标度

（a）　优化模型生成的三类网络：星型网络、无标度网络和指数网络。处于未标记区域的网络，其拓扑结构类型未知。

星型网络所在区域的右侧竖直边界是$\delta=(1/2)^{1/2}$。这是模型所定义的在单位正方形上两个节点的最大距离的倒数。当$\delta<(1/2)^{1/2}$时，对于任意新节点都有$\delta d_{ij}<1$；根据公式**4.28**，新节点连接到中心节点的成本$C_i=\delta d_{ij}+0$，总是小于新节点连接到其他任何节点的成本$f(i,j)=\delta d_{ij}+1$。因此，当$\delta<(1/2)^{1/2}$时，所有节点都会连接到节点0，形成一个由单个枢纽节点主导的网络（即图c的星型辐射网络）。

无标度区域的斜向边界为$\delta=N^{1/2}$。实际上，如果节点随机放置在单位正方形中，相邻节点之间的典型距离是$N^{-1/2}$。因此，如果$d_{ij}\sim N^{-1/2}$，对于大部分节点对，有δd_{ij}

$\geq h_j$ 。通常来说，到达中心节点的路径长度h_j 增长得比N 慢（小世界网络中$h_j \sim \log N$ ；无标度网络中$h_j \sim \ln\ln N$ ）。因此，C_i 主要由δd_{ij} 项决定，最小的C_i 可以通过最小化这个距离依赖项得到。注意，严格来说，这种情况只在$N \to \infty$的极限下才能发生。在白色区域，我们没有度分布的解析形式解。

（**b**） 图a中的三种网络的度分布（$N = 10^4$ ）。

（**c**） 不同δ 值下优化模型生成的典型拓扑。节点大小正比于节点的度。

（**d**） 使用**4.7**节 介绍的方法测量偏好连接函数。我们从具有$N = 10\ 000$个节点的网络开始，逐个增加新节点并测量其连接到的节点的度，重复该过程10 000次得到偏好连接函数$\Pi(k)$。从图中可以看出，无标度网络中存在线性偏好连接，而星型网络和指数网络中没有线性偏好连接。

总之，我们可以构造出在定义中没有显式偏好连接函数$\Pi(k)$、却能生成无标度网络的模型。这些模型通过隐含的偏好连接来生成无标度网络。产生偏好连接的微观机制有两种截然不同的起源（图**4-17**）：一种源于随机过程，例如链接选择机制或复制机制；另一种源于优化过程，新节点在选择连接对象时在一些相互冲突的目标中寻求平衡。注意，本节讨论的这些机制都能形成巴拉巴西－阿尔伯特模型中的线性偏好连接。截至目前，我们还没有发现哪种机制能够产生**4.8**节 中讨论的非线性偏好连接。

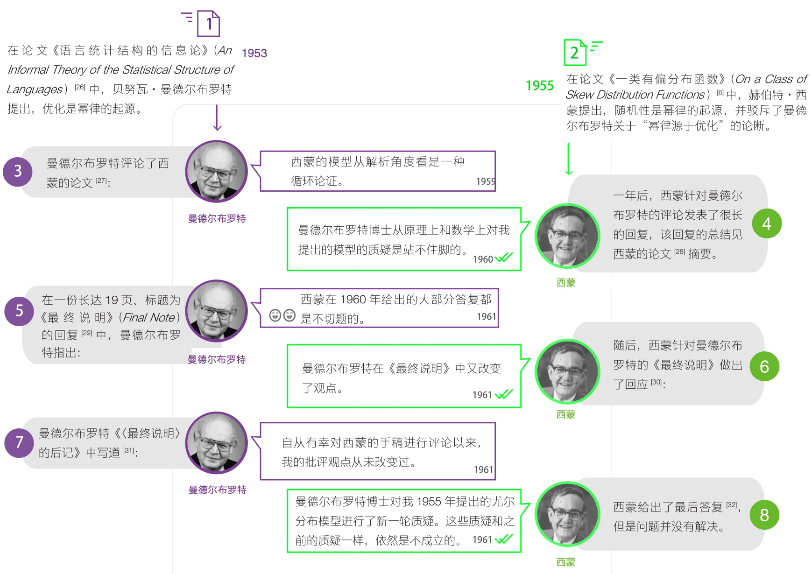


图4-17　偶然还是必然：一个古老的争论

幂律的两种明显对立的解释——随机和优化之间的争论持续已久：在20世纪60年代，赫伯特·西蒙和贝努瓦·曼德尔布罗特曾公开辩论过此问题。西蒙认为，词频的幂律分布源自偏好连接；而曼德尔布罗特坚持基于优化的框架。这次争论持续了两年，涉及7篇论文，是有记录以来最激烈的学术争论之一。

如今，在网络背景下，西蒙的观点似乎占了上风：复杂网络中观察到的幂律是由随机性和偏好连接所致。然而，曼德尔布罗特提出的基于优化的方法在解释偏好连接的起源方面起着重要作用。因此，最终看来他们俩都是对的。

本节讨论的机制多样性表明，线性偏好连接是多么常见，它会出现在很多完全不同的系统中。线性偏好连接既可以源自理性选择，也可以源自随机行为 [25] 。而大多数复杂系统是由这两种机制共同作用形成的。因此，无论是从运气角度来看，还是从理性角度来看，偏好连接都是最终的赢家。

4.10 直径与集聚系数

为了更完整地刻画巴拉巴西 - 阿尔伯特模型，本节我们将讨论这个模型的网络直径和集聚系数。

直径

网络直径表示网络中节点间的最大距离。对于巴拉巴西 - 阿尔伯特网络而言，当$m>1$并且N较大时，网络直径为 [33] ，[34] ：

$$\langle d\rangle \sim \frac{\ln N}{\ln\ln N} \tag{4.29}$$

因此，网络直径的增长要比$\ln N$ 慢。也就是说，巴拉巴西 - 阿尔伯特网络的直径要比同等大小的随机网络的直径小，当N 较大时，这一差异更加明显。

尽管公式**4.29** 是关于网络直径的表达式，网络平均距离$\langle d\rangle$ 的情形也与之类似。实际上，如图**4-18** 所示，当N较小时，$\ln N$刻画了网

络平均距离$\langle d\rangle$ 关于N 的变化情况；然而，当N 较大时（$N \geq 10^4$ ），对数修正项$\ln\ln N$ 的影响就不能忽略了。

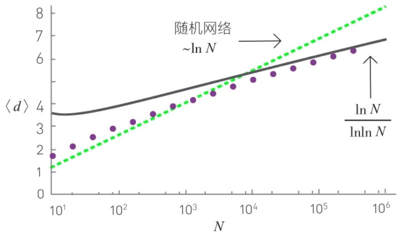


图4-18　平均距离

巴拉巴西 - 阿尔伯特模型中平均距离和网络大小之间的关系。实线对应公式**4.29** 给出的结果，虚线对应随机网络上的结果——公式**2.19** 。由于公式给出的解析解不能精确地给出依赖关系的系数，因此这些线条并不表示拟合，只是用来刻画平均距离关于N 的变化趋势。图中给出的数据点是在$m=2$时独立运行10次取平均值得到的。

集聚系数

巴拉巴西 - 阿尔伯特网络中的集聚系数为（进阶阅读**4.C** ）[35]，[36]：

$$\langle C\rangle \sim \frac{(\ln N)^2}{N} \qquad (4.30)$$

公式**4.30** 和随机网络中集聚系数对$1/N$ 的依赖关系（图**4-19** ）大不相同。差异主要来自$(\ln N)^2$ ，这一项在N 较大时大大增加了集聚系数。因此，巴拉巴西 - 阿尔伯特网络中的集聚系数比随机网络中的集聚系数要高。

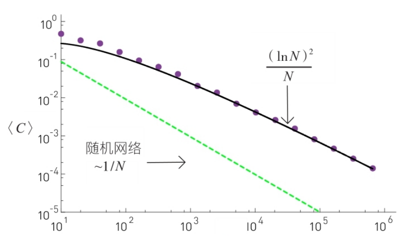


图4-19　集聚系数

巴拉巴西 - 阿尔伯特模型中平均集聚系数和网络大小之间的关系。实线对应公式**4.30** 给出的解析解，虚线对应随机网络上的结果——$\langle C \rangle \sim 1/N$。图中给出的数据点，是在$m=2$时独立运行10次取平均值得到的。虚线和实线并不表示拟合，只是用来刻画平均距离关于N的变化趋势。

4.11
小结

巴拉巴西 - 阿尔伯特模型给我们带来的最重要的启示是，网络的结构和演化密不可分。实际上，在埃尔德什 - 雷尼模型、瓦茨 - 斯托加茨模型、配置模型和隐参数模型等模型中，建模者的作用体现在如何在固定数量的节点间放置链接。回顾一下我们之前给出的比喻，这些模型生成的网络和真实网络之间的关系就像一幅画的照片和这幅画本身之间的关系一样：照片可能看起来和真实的画很像，但是生成照片的过程和创作原画的过程完全不同。巴拉巴西 - 阿尔伯特模型的初衷在于刻画网络从头开始逐渐成长的过程。因此，该模型是在重新作画，尽可能画得和原来一样。巴拉巴西 - 阿尔伯特模型的建模思想很简单：要理解复杂系统的拓扑结构，我们需要描述它是如何形成的。

接下来，在探索某种网络性质和我们的预期偏离程度如何时，随机网络模型、配置模型和隐参数模型还将继续发挥重要作用。不过，如果我们打算解释某个网络性质的起源，就不得不使用巴拉巴西 - 阿尔伯特模型这样的能够刻画系统起源的模型了。

巴拉巴西 - 阿尔伯特模型引发了一个基本问题：生长和偏好连接的组合是网络具有无标度特性的原因吗？关于这个问题，本节给出了一个充分且必要的论断。首先，生长和偏好连接对于生成无标度网络都是必要的，若二者缺一，无标度性质或者稳定的无标度性质就会消失。其次，生长和偏好连接机制都存在时，的确能够生成无标度网络。不过，这样的论断还是留下了一个开放性问题：这两种机制能够解释所有网络的无标度性质吗？有些真实的无标度性质是否可能是由其他完全不同的机制带来的呢？答案参见**4.10**节　　，在那里我们介绍

了链接选择模型、复制模型和优化模型，这些模型的确没有内置的偏好连接机制，却能产生无标度网络。这些模型之所以能够产生无标度网络，是因为它们可以隐含地形成线性偏好连接函数$\Pi(k)$。这一发现突出了一个更一般性的规律：迄今为止，我们在所有已知的具有无标度性质的模型和真实系统中都发现了偏好连接。因此，巴拉巴西 - 阿尔伯特模型的基本机制看上去的确刻画了无标度拓扑的起源。

然而，真实系统的很多性质是巴拉巴西 - 阿尔伯特模型无法描述的：

- 该模型预测度指数为$\gamma=3$，而真实网络的度指数可以在2到5之间取值（表**3-2**）。
- 万维网、引文网络等许多真实网络都是有向的，而巴拉巴西 - 阿尔伯特模型生成的网络是无向网络。
- 网络中观测到的很多过程，例如已有节点之间形成链接、链接消失、节点消失等，在巴拉巴西 - 阿尔伯特模型中都没有。
- 该模型无法区分节点的一些内在特性，比如一篇论文的创新性或一个网页的有用性。
- 虽然巴拉巴西 - 阿尔伯特模型偶尔会被用来建模互联网或细胞，但实际上它并不是针对任何特定真实网络的细节而设计的。该模型是一个最简化的、原理性的模型，其主要初衷是刻画无标度特性形成的基本机制。因此，如果我们想理解互联网、细胞或万维网等系统的演化，还需要考虑在这些系统演化过程中起重要作用的细节，例如万维网的有向性、已有节点之间形成内部链接的可能性，以及节点和链接的消失。

在第**5**章 中我们将看到，这些局限性都可以被系统地解决。

边栏4.4

快速回顾：

巴拉巴西 - 阿尔伯特模型

节点数

$$N = t$$

链接数

$$N = mt$$

平均度

$$\langle k \rangle = 2m$$

度动力学

$$k_i(t) = m = (t/t_i)^{\beta}$$

动态指数

$$\beta = 1/2$$

度分布

$$p_k \sim k^{-\gamma}$$

度指数

$$\gamma = 3$$

平均度

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N}$$

集聚系数

$$\langle C \rangle \sim \frac{(\ln N)^2}{N}$$

4.12

课后习题

习题1　生成巴拉巴西 - 阿尔伯特网络

借助计算机，用巴拉巴西 - 阿尔伯特模型生成一个$m=4$、$N=10^4$ 的网络。初始状态是一个大小为$m_0=4$的完全图。

（**1**）在网络大小达到10^2，10^3，10^4 的中间步骤，测量网络的度分布。

（**2**）相应地，将这些中间步骤的度分布绘制在一起进行对比，并用幂律分布进行拟合，得到度指数γ。看看网络生成过程中所得的度分布是否“收敛”？

（**3**）绘出上述中间步骤的累积度分布。

（**4**）测量平均集聚系数关于N 的函数。

（**5**）参照图**4-6a**，测量一些节点的度动态变化过程。这些节点分别是：某一个初始节点、$t=100$时加入网络的节点、$t=1\ 000$时加入网络的节点和$t=5\ 000$时加入网络的节点。

习题2　有向的巴拉巴西 - 阿尔伯特模型

考虑巴拉巴西 - 阿尔伯特模型的一个变种，在每个时间步新增一个节点，然后从已有节点中选择一个节点，添加一条从新节点指向该节点的有向链接。选择到节点i 的概率为：

$$\Pi\left(k_i^{\text{in}}\right)=\frac{k_i^{\text{in}}+A}{\sum_j\left(k_j^{\text{in}}+A\right)}$$

这里，k_i^{in} 表示节点i 的入度，常数A 对于所有节点都是一样的。每个新节点有m 条有向链接。

（**1**）使用速率方程法计算出网络的入度分布和出度分布。

（**2**）利用伽马函数和贝塔函数的性质，你能找到入度分布的幂指数吗？

（3） 当$A=0$时，入度分布的幂指数和标准巴拉巴西-阿尔伯特模型的幂指数$\gamma=3$取值不同。为什么？

习题3 复制模型

使用速率方程法证明，有向复制模型能够生成入度度指数为
$\gamma_{\text{in}}=\frac{2-p}{1-p}$ 的无标度网络。

习题4 无偏好连接的增长

模型A中，网络生长过程中新节点随机选择m 个已经存在的节点并与之相连。请推导出模型A的度分布——公式**4.18** 。借助计算机，使用模型A生成一个大小为10^4 的网络，测量其度分布并检查其是否与公式**4.18** 吻合。

4.13 进阶阅读4.A 推导度分布

有许多解析方法可以计算度指数的精确形式——公式**4.11** 。接下来，我们使用速率方程法[12]，[13]推导度指数的精确形式。该方法非常普适，可以用于分析很多生长网络的性质。因此，这里介绍的计算过程对于很多系统都有直接的意义，包括万维网建模的相关模型[16]，[17]，[18]和通过基因复制描述蛋白质相互作用的网络演化的模型[19]，[20]。

我们用$N(k,t)$表示时间为t时度为k的节点数。度分布$p_k(t)$和$N(k,t)$的关系为$p_k(t)=N(k,t)/N(t)$。由于我们在每个时间步向网络中添加一个新节点，因此有$N=t$ 。也就是说，在任意时刻，节点总数都等于时间步数（边栏**4.2** ）。

我们把偏好连接写作如下形式：

$$\Pi(k)=\frac{k}{\sum_j k_j}=\frac{k}{2mt} \quad (4.31)$$

其中，$2m$ 体现了这样一个事实：无向网络中每条链接对两个节点的度有贡献。我们的目标是计算出网络中新增一个节点后，度为k的节点数的变化。为此，我们检查新节点加入后改变$N(k,t)$和$p_k(t)$的两种情况：

（1）新节点连接到了度为k 的节点，使它的度变成了$(k+1)$，从而减小了$N(k,t)$。

（2）新节点连接到了度为$(k-1)$的节点，使它的度变成了k，从而增大了$N(k,t)$。

新节点加入后，连接到度为k 节点的链接数为：

$$\frac{k}{2mt}\times Np_k(t)\times m=\frac{k}{2}p_k(t) \quad (4.32)$$

公式**4.32** 中左侧第一项刻画了新节点连接到度为k 的节点的概率（偏好连接）；第二项是度为k 的节点数：度为k 的节点数越多，新节点连接到一个度为k 的节点的概率越大；第三项是新节点的度：m越大，新节点连接到度为k 的节点的机会越多。接下来，我们把公式**4.32** 应用于前面所述的（1）和（2）两种情况：

（1）原来度为k、后来因获得一条新链接而变成度为$(k+1)$的节点数量为：

$$\frac{k}{2}p_k(t) \quad (4.33)$$

（2）原来度为$(k-1)$、后来因获得一条新链接而变成度为k 的节点数量为：

$$\frac{k-1}{2}p_{k-1}(t) \quad (4.34)$$

合并公式**4.33** 和公式**4.34**，可得到增加新节点后度为k 的节点的期望数量：

$$(N+1)p_k(t+1)=Np_k(t)+\frac{k-1}{2}p_{k-1}(t)-\frac{k}{2}p_k(t) \qquad (4.35)$$

该公式适用于所有度$k>m$的节点。由于网络中没有度为$k=0,1,\ldots,m-1$的节点（每个新节点的度为m），对于度为m的节点，我们还需要一个单独的公式。类似于公式**4.35**的推导过程，我们有：

$$(N+1)p_m(t+1)=Np_m(t)+1-\frac{m}{2}p_m(t) \qquad (4.36)$$

公式**4.35**和公式**4.36**是递归计算p_k的第一步。进而，由于我们要寻找的是一个稳定度分布——参照图**4-6**中的数值模拟，在$N=t\rightarrow\infty$的极限下有$p_k(\infty)=p_k$。据此，我们可以把公式**4.35**和公式**4.36**的左侧写为：

$$(N+1)p_k(t+1)-Np_k(t)\rightarrow Np_k(\infty)+p_k(\infty)-Np_k(\infty)=p_k(\infty)=p_k$$

$$(N+1)p_m(t+1)-Np_m(t)\rightarrow p_m$$

因此，速率方程**4.35**和**4.36**的形式为：

$$p_k=\frac{k-1}{k+2}p_{k-1}\quad k>m \qquad (4.37)$$

$$p_m=\frac{2}{m+2} \qquad (4.38)$$

注意，通过变量替换$k\rightarrow k+1$，公式**4.37**可以重写为：

$$p_{k+1}=\frac{k}{k+3}p_k \qquad (4.39)$$

使用递推方法可以求得度分布。也就是说，我们先用公式**4.38**写出最小度$k=m$的度分布，然后利用公式**4.39**递推计算出更大度时的p_k：

$$p_{m+1}=\frac{m}{m+3}p_m=\frac{2m}{(m+2)(m+3)}$$

$$p_{m+2}=\frac{m+1}{m+4}p_{m+1}=\frac{2m(m+1)}{(m+2)(m+3)(m+4)} \quad (4.40)$$

$$p_{m+3}=\frac{m+2}{m+5}p_{m+2}=\frac{2m(m+1)}{(m+3)(m+4)(m+5)}$$

至此，我们得到了一个简单的递归形式：将分母中的$m+3$替换为k，便得到了观察到度为k的节点的概率：

$$p_k=\frac{2m(m+1)}{k(k+1)(k+2)} \quad (4.41)$$

这便是巴拉巴西 - 阿尔伯特模型度分布的精确形式。

注意：

- 当k较大时，公式**4.41**变成了$p_k \sim k^{-3}$，和数值模拟的结果一致。
- 公式**4.11**和**4.41**的前置因子和公式**4.9**的前置因子不同。
- 这一形式是在《物理评论快报》曾发表的两篇论文[12]，[13]中独立推导出的，其正确性的严格数学证明请参考波罗巴斯的论文[10]。

最后，这个速率方程提供了一个满足度分布的连续介质方程[16]。从速率方程

$$p_k=\frac{k-1}{2}p_{k-1}-\frac{k}{2}p_k \quad (4.42)$$

出发，我们可以写出：

$$2p_k=(k-1)p_{k-1}-kp_k=-p_{k-1}-k\left[p_k-p_{k-1}\right] \quad (4.43)$$

$$2p_k=-p_{k-1}-k\frac{p_k-p_{k-1}}{k-(k-1)}\approx -p_{k-1}-k\frac{\partial p_k}{\partial k} \quad (4.44)$$

进而得到：

$$p_k = -\frac{1}{2}\frac{\partial[kp_k]}{\partial k} \quad (4.45)$$

通过检验，可以看出方程**4.45** 的解是：

$$p_k \sim k^{-3} \quad (4.46)$$

4.14 进阶阅读4.B 非线性偏好连接

这一节中，我们将推导由偏好连接公式**4.22** 支配的非线性巴拉巴西 - 阿尔伯特模型的度分布。我们参照参考文献[13]的方式，但是调整了计算过程以覆盖$m>1$的情况。

严格来说，公式**4.22** 所描述的稳定度分布只在$\alpha \leq 1$时存在。如**4.8**节 中所解释的那样，当$\alpha >1$时，少数节点吸引了固定比例的链接，p_k 不再与时间无关。因此，接下来我们只讨论$\alpha \leq 1$时情况。

在非线性巴拉巴西 - 阿尔伯特模型中，每个时间步新增加一个节点，新节点带来m 条新链接。每条新链接将新节点和网络中已经存在的一个节点相连，新节点连接到度为k_i 的节点的概率为：

$$\Pi(k_i) = \frac{k_i^{\alpha}}{M(\alpha,t)} \quad (4.47)$$

这里$0<\alpha \leq 1$，并且

$$M(\alpha,t) = t\sum_k k^{\alpha} p_k(t) = t\mu(\alpha,t) \quad (4.48)$$

是归一化因子，$t = N(t)$表示节点数。

注意，

$$\mu(0,t)=\sum_k p_k(t)=1\text{，}\quad \mu(1,t)=\sum_k kp_k(t)=\langle k\rangle=2mt/N$$

是平均度。由于$0<\alpha\leq 1$，有

$$\mu(0,t)\leqslant\mu(\alpha,t)\leqslant\mu(1,t) \tag{4.49}$$

因此，当时间趋于无穷大时

$$\mu(\alpha,t\to\infty)=\text{常数} \tag{4.50}$$

这个常数的精确值我们稍后计算。简单起见，我们记$\mu=\mu(\alpha,t\to\infty)$。

根据进阶阅读**4.A** 介绍中介绍的速率方程法，网络度分布的速率方程可以写为：

$$(t+1)p_k(t+1)=tp_k(t)+\frac{m}{\mu(\alpha,t)}$$

$$\left[(k-1)^{\alpha}p_{k-1}(t)-k^{\alpha}p_k(t)\right]+\delta_{k,m} \tag{4.51}$$

等式右侧第一项描述了度为$(k-1)$的节点获得新链接的速率；第二项描述了度为k的节点的流失速度——因为获得新的链接变成了度为$(k+1)$的节点；最后一项表示新加入的度为m的节点。

当$t\to\infty$时，我们近似有$p_k=p_k(t+1)=p_k(t)$。将$k=m$代入公式**4.51**得到：

$$\begin{aligned}p_m&=-\frac{m}{\mu}-m^{\alpha}p_m+1\\p_m&=-\frac{\mu/m}{\mu/m+m^{\alpha}}\end{aligned} \tag{4.52}$$

当$k>m$时，

$$p_k=\frac{m}{\mu}\left[(k-1)^{\alpha}p_{k-1}-k^{\alpha}p_k\right] \tag{4.53}$$

$$p_k=\frac{(k-1)^{\alpha}}{\mu/m+k^{\alpha}}p_{k-1} \tag{4.54}$$

递归求解方程4.53 可得：

$$p_m = \frac{\mu / m}{\mu / m + m^{\alpha}} \quad (4.55)$$

$$p_{m+1} = \frac{m^{\alpha}}{\mu / m + (m+1)} \frac{\mu / m}{\mu / m + m^{\alpha}} \quad (4.56)$$

$$p_k = \frac{\mu / m}{k^{\alpha}} \prod_{j=m}^{k} \left(1 + \frac{\mu / m}{j^{\alpha}}\right)^{-1} \quad (4.57)$$

为了确定k 较大时p_k 的情况，我们对公式4.57 取对数：

$$\ln p_k = \ln(\mu / m) - \alpha \ln k - \sum_{j=m}^{k} \left(1 + \frac{\mu / m}{j^{\alpha}}\right) \quad (4.58)$$

使用级数展开$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} / n \bullet x^n$，我们得到：

$$\ln p_k = \ln(\mu / m) - \alpha \ln k - \sum_{j=m}^{k} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (\mu / m)^n j^{-n\alpha} \quad (4.59)$$

此处对j 的求和可以近似为积分

$$\sum_{j=m}^{k} j_X^{-n\alpha} \approx \int_m^k x^{-n\alpha} dx = \frac{1}{1 - n\alpha} \left(k^{1-n\alpha} - m^{1-n\alpha}\right) \quad (4.60)$$

在$n\alpha = 1$的特例下，上式变为

$$\sum_{j=m}^{k} j_X^{-1} \approx \int_m^k x^{-1} dx = \ln k - \ln m \quad (4.61)$$

因此，我们有

$$\ln p_k = \ln(\mu / m) - \alpha \ln k - \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \frac{(\mu / m)^n}{1 - n\alpha} \left(k^{1-n\alpha} - m^{1-n\alpha}\right) \quad (4.62)$$

因此，度分布具有如下形式：

$$p_k = C_\alpha k^{-\alpha} e^{-\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \frac{(\mu/m)^n}{1-n\alpha} k^{1-n\alpha}} \quad (4.63)$$

其中

$$C_\alpha = \frac{\mu}{m} e^{\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \frac{(\mu/m)^n}{1-n\alpha} m^{1-n\alpha}} \quad (4.64)$$

指数中的衰减项并不影响$k \to \infty$时的渐进行为，仅当$1-n\alpha \geq 1$时才起作用。因此，p_k和α的依赖关系为：

$$p_k \sim f(x) = \begin{cases} k^{-\alpha} e^{\frac{-\mu/m}{1-\alpha} k^{1-\alpha}} & 1/2 < \alpha < 1 \\ k^{-\frac{1}{2}+\frac{1}{2}\left(\frac{\mu}{m}\right)^2} e^{-\frac{1}{2}\frac{\mu}{m} k^{-2}} & \alpha = 1/2 \\ k^{-\alpha} e^{\frac{-\mu/m}{1-\alpha} k^{1-\alpha} + \frac{1}{2}\frac{(\mu/m)^2}{1-2\alpha} k^{1-2\alpha}} & 1/3 < \alpha < 1/2 \\ \ldots\ldots & \end{cases} \quad (4.65)$$

也就是说，对于$1/2<\alpha<1$，度分布服从广延指数分布。当我们减小α时，每当α的取值变得比$1/n$小（n是一个整数），就会有一个新的修正项开始起作用。

当$\alpha \to 1$时，度分布关于k的函数为k^{-3}，和巴拉巴西 - 阿尔伯特模型一样。实际上，当$\alpha=2$时，我们有$\mu=2$，且

$$\lim_{\alpha \to 1} \frac{k^{1-\alpha}}{1-\alpha} = \ln k \quad (4.66)$$

因此，$p_k \sim k^{-1} \exp(-2\ln k) = k^{-3}$。

最后，我们计算$\mu(\alpha) = \sum_j j^\alpha p_j$。据此，公式4.58 的求和项可以写为

$$\sum_{k=m}^{\infty} k^\alpha p_k = \sum_{k=m}^{\infty} \frac{\mu(\alpha)}{m} \prod_{j=m}^{k} \left(1 + \frac{\mu(\alpha)/m}{j^\alpha}\right)^{-1} \quad (4.67)$$

$$1=\frac{1}{m}\sum_{k=m}^{\infty}\prod_{j=m}^{k}\left(1+\frac{\mu(\alpha)/m}{j^{\alpha}}\right)^{-1} \qquad (4.68)$$

我们通过数值求解方程**4.68** 得到$\mu(\alpha)$。

4.15 进阶阅读4.C 集聚系数

这一节中，我们将推导巴拉巴西 - 阿尔伯特模型的集聚系数——公式**4.30** 。推导过程遵循克莱姆（Klemm）和埃格鲁兹(V. M. Eguiluz)提出的观点[35]，并有波罗巴斯的精确计算作为支持[36]。

我们的目标是计算模型中三角形的期望数量，这个量和集聚系数密切相关（**1.10**节）。我们用$P(i,j)$表示节点i和节点j之间存在链接的概率。因此，三个节点i，j，l组成三角形的概率为$P(i,j)P(i,l)P(j,l)$。度为k_l的节点l参与的三角形期望数量是节点l与网络中任意节点i和j组成三角形的概率之和。假设度的取值是连续的，近似得到：

$$Nr_1(\Delta)=\int_{i=1}^{N}di\int_{j=1}^{N}dj\,P(i,j)P(i,l)P(j,l) \qquad (4.69)$$

接下来，我们需要计算$P(i,j)$，这需要考虑巴拉巴西 - 阿尔伯特模型是如何演化的。由于在每个时间步我们只添加一个新节点（边栏**4.2** 的事件时间），我们可以用$t_j=j$表示节点j的加入时间。因此，节点j加入网络时连接到度为k_i的节点i的概率可以表述为偏好连接

$$P(i,j)=m\Pi(k_i(j))=m\frac{k_i(j)}{\sum_{l=1}^{j}k_l(j)}=m\frac{k_i(j)}{2mj} \qquad (4.70)$$

根据公式**4.7** ，我们可以写出

$$k_i(t)=m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}=m\left(\frac{j}{i}\right)^{\frac{1}{2}} \qquad (4.71)$$

这里我们利用了节点j 的到达时间$t_j=j$ 以及节点i 的到达时间$t_i=i$。因此，公式4.70 变成了

$$P(i,j)=\frac{m}{2}(ij)^{-\frac{1}{2}} \quad (4.72)$$

根据这一结果，我们计算公式4.69 中的三角形数量，即

$$Nr_1(\Delta)=\int_{i=1}^{N}di\int_{j=1}^{N}djP(i,j)P(i,l)P(j,l) \quad (4.73)$$
$$=\frac{m^3}{8}\int_{i=1}^{N}di\int_{j=1}^{N}dj(ij)^{-\frac{1}{2}}(il)^{-\frac{1}{2}}(jl)^{-\frac{1}{2}}$$
$$=\frac{m^3}{8l}\int_{i=1}^{N}\frac{di}{i}\int_{j=1}^{N}\frac{dj}{j}=\frac{m^3}{8l}(\ln N)^2$$

集聚系数可以写作：

$$C_l=\frac{2Nr_1(\Delta)}{k_l(k_l-1)}$$

因此，我们有：

$$C_l=\frac{\frac{m^3}{4l}(\ln N)^2}{k_l(N)(k_l(N)-1)} \quad (4.74)$$

为了简化公式4.74 ，我们根据公式4.7 得到

$$k_l(N)=m\left(\frac{N}{l}\right)^{\frac{1}{2}} \quad (4.75)$$

这就是节点l 在时间$t=N$ 时的度。因此，当k_l 较大时，我们有：

$$k_l(N)(k_l(N-1))\approx k_l^2(N)=m^2\frac{N}{l} \quad (4.76)$$

因而，我们可以写出巴拉巴西 - 阿尔伯特模型的集聚系数

$$C_l=\frac{m}{4}\frac{(\ln N)^2}{N} \quad (4.77)$$

可以看出C_l和l无关，因此我们得到了公式4.30的结果。

4.16 参考文献

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[2] F. Eggenberger and G. Pólya. Über die Statistik Verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik* , 3:279-289, 1923.

[3] G.U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. Philosophical Transactions of the Royal Society of London. *Series B* , 213:21-87, 1925.

[4] R. Gibrat. *Les Inégalités économiques* . Paris, France, 1931.

[5] G. K. Zipf. *Human behavior and the principle of least resort* . Addison-Wesley Press, Oxford, England, 1949.

[6] H. A. Simon. On a class of skew distribution functions. *Biometrika* , 42:425-440, 1955.

[7] D. De Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* , 27:292-306, 1976.

[8] R. K. Merton. The Matthew effect in science. *Science* , 159:56-63, 1968.

[9] A.-L. Barabási. *Linked: The new science of networks* .Perseus, New York, 2002.

[10] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms* , 18:279-290, 2001.

[11] A.-L. Barabási, H. Jeong, R. Albert. Mean-field theory for scale-free random networks. *Physica A* , 272:173-187, 1999.

[12] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett* ., 85:4633-4636, 2000.

[13] P.L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett* ., 85:4629-4632, 2000.

[14] H. Jeong, Z. Néda. A.-L. Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters* , 61:567-572, 2003.

[15] M.E.J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E* , 64:025102, 2001.

[16] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of networks* . Oxford Clarendon Press, 2002.

[17] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: measurements, models and methods. Proceedings of the International Conference on Combinatorics and Computing, 1999.

[18] R. Kumar, P. Raghavan, S. Rajalopagan, D. Divakumar,A.S. Tomkins, and E. Upfal. The Web as a graph. Proceedings of the 19th Symposium on principles of database systems, 2000.

[19] R. Pastor-Satorras, E. Smith, and R. Sole. Evolving protein minteraction networks through gene duplication. *J.Theor. Biol* . 222:199-210, 2003.

[20] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs* , 1:38-44, 2003.

[21] G.S. Becker. *The economic approach to Human Behavior* .

Chicago, 1976.

[22] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically optimized trade-offs: a new paradigm for power laws in the internet. In Proceedings of the 29th International Colloquium on Automata, Languages, and Programming (ICALP),pages 110-122, Malaga, Spain, July 2002.

[23] R.M. D'Souza, C. Borgs, J.T. Chayes, N. Berger, and R.D.Kleinberg. Emergence of tempered preferential attachment from optimization. *PNAS* , 104, 6112-6117, 2007.

[24] F. Papadopoulos, M. Kitsak, M. Angeles Serrano, M.Boguna, and D. Krioukov. Popularity versus similarity in growing networks. *Nature* , 489: 537, 2012.

[25] A.-L. Barabási. Network science: luck or reason. *Nature* 489: 1-2, 2012.

[26] B. Mandelbrot. An Informational Theory of the Statistical Structure of Languages. In *Communication Theory* ,edited by W. Jackson, pp. 486-502. Woburn, MA: Butterworth, 1953.

[27] B. Mandelbrot. A note on a class of skew distribution function: analysis and critique of a Paper by H.A. Simon. *Information and Control* , 2: 90-99, 1959.

[28] H.A. Simon. Some Further Notes on a Class of Skew Distribution Functions. *Information and Control* , 3: 80-88, 1960.

[29] B. Mandelbrot. Final Note on a Class of Skew Distribution Func tions: Analysis and Critique of a Model due to H.A. Simon. *Information and Control* , 4: 198-216, 1961.

[30] H.A. Simon. Reply to final note. *Information and Control* , 4: 217-223, 1961.

[31] B. Mandelbrot. Post scriptum to final note. *Information and Control* , 4: 300-304, 1961.

[32] H.A. Simon. Reply to Dr. Mandelbrot's Post Scriptum. *Information and Control* , 4: 305-308, 1961.

[33] R. Cohen and S. Havlin. Scale-free networks are ultra small. *Phys. Rev. Lett* ., 90:058701, 2003.

[34] B. Bollobás and O.M. Riordan. The diameter of a scale-free random graph. *Combinatorica* , 24:5-34, 2004.

[35] K. Klemm and V.M. Eguluz. Growing scale-free networks with small-world behavior. *Phys. Rev. E* , 65:057102, 2002.

[36] B. Bollobás and O.M. Riordan. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks* ,edited by S. Bormholdt and A. G. Schuster, Wiley, 2003.

第5章 演化网络

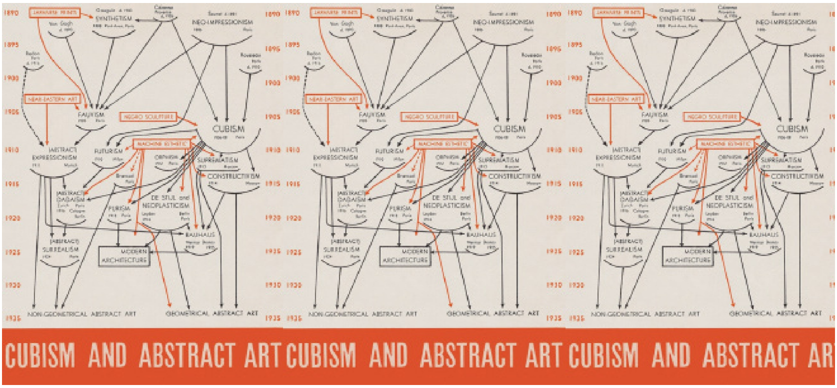


图**5-0** 阿尔弗雷德·巴尔的立体主义和抽象艺术

阿尔弗雷德·巴尔（Alfred Barr，1902—1981），美国艺术史学家，也是纽约现代艺术博物馆（MoMA）的首任馆长，现代艺术发展历程中最具影响力的推动者之一。巴尔在1936年绘制的这幅作品，使用了网络框架来表达现代艺术循环往复的发展历程。在MoMA首次展览立体主义和抽象艺术运动时，这幅作品出现在展览目录的封面上。

本章索引

5.1 本章概览 / **247**

5.2 比安科尼 - 巴拉巴西模型 / **248**

5.3 测量适应性 / **252**

5.4 玻色 - 爱因斯坦凝聚 / **257**

5.5 演化网络 / **262**

5.6 小结 / **271**

5.7 课后习题 / **275**

5.8 进阶阅读**5.A**
比安科尼 - 巴拉巴西模型的解析解 / **277**

5.9 参考文献 / **280**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

罗伯塔·西纳特拉 **ROBERTA SINATRA**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

萨拉·莫里森 **SARAH MORRISON**

莫罗·马蒂诺 **MAURO MARTINO**

阿迈勒·侯赛尼 **AMAL HUSSEINI**

菲利普·赫费尔 **PHILIPP HOEVEL**

妮科尔·萨迈 **NICOLE SAMAY**

5.1 本章概览

谷歌在万维网诞生6年后创立，是网络搜索的后来者。在20世纪90年代后期，两家起步较早的搜索引擎公司Alta Vista和Inktomi统治着搜索市场。然而，第三个进入市场的谷歌不仅很快成为领先的搜索引擎，还以令人难以置信的速度得到大量其他网站的链接，从而在2000年成为万维网中最大的枢纽节点 [1] 。可是这种状态并未持续太久：2001年，一家比谷歌更年轻的公司——Facebook取而代之，成为万维网中最大的节点。

万维网的竞争局面凸显出我们的模型框架的一项重要不足：迄今为止，我们遇到的所有网络模型均不能建模这种后来居上的竞争局面。事实上，埃尔德什－雷尼模型中哪个节点会成为最大节点完全是随机的。巴拉巴西－阿尔伯特模型给出了更接近现实的描述，指出每个节点的度的增长服从$k(t) \sim t^{1/2}$ 。这意味着最老的节点总是有最多的链接，也就是商业领域常说的先发优势。这也意味着后来者永远不可能成为最大的枢纽节点。

在现实中，一个节点的增长速度并不完全取决于其年龄。网页、公司、演员的内在品质影响了其获得链接的速度。有一些节点虽然出现得较晚，却能在短时间内获取大量链接；另一些出现得较早，却从未获得大量链接。本章的目标是理解节点获取链接的能力如何影响网络拓扑。除了万维网的竞争局面，我们也将探索诸如删除节点和链接（图**5-1**）、节点衰老（真实网络中经常观察到这一现象）等其他过程如何改变网络的演化和拓扑。我们的目标是发展出一套自洽的、可以随意调节的演化网络理论，以预测多种真实网络的动态过程和拓扑。



图**5-1**　时装区

时装区位于纽约曼哈顿第五大道和第九大道之间，横跨34街到42街。自20世纪初以来，时装区一直是美国时尚界的中心。该地区的核心位置树立着两座雕塑，以向其过往致敬，这两座雕塑分别是穿过纽扣的针和犹太裁缝。

纽约的时装业是网络衰退的一个典型例子，可以帮助我们理解节点流失是如何重塑网络拓扑的（边栏**5.5**　）。揭示节点和链接的流失过程对网络拓扑的影响，是本章的目标之一。

5.2
比安科尼 - 巴拉巴西模型

有些人善于将萍水相逢的关系转变为长久的社交关系；有些公司能够把每一个客户变成忠实的合作伙伴；有些网页能吸引访问者频频回顾。这些成功的节点的共性是：它们具有某种使其脱颖而出的内在属性。我们把这种内在属性称为适应性。

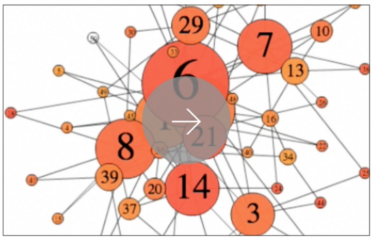
适应性是一个人将萍水相逢转变为长久友谊的天赋；适应性是一家公司在竞争中赢得客户的诀窍；适应性是一个网页与众多网页竞争、获取人们关注的能力。对于个人，适应性也许与基因有关；对于公司，适应性也许与创新能力和管理水平有关；对于网站，适应性也许取决于网站的内容。

在巴拉巴西 - 阿尔伯特模型中，我们假设一个节点的成长速度完全由该节点的度决定。为了引入节点的适应性，我们假设偏好连接取决于节点的适应性η　和节点度k　的乘积。由此而得到的模型被称为比

安科尼 - 巴拉巴西模型，也被称为适应性模型。该模型包含以下两个步骤[2]，[3]：

1.生长

在每个时间步，网络中加入一个新节点j，该节点具有m条链接和适应性η_j。其中η_j是从适应性分布ρ（η）中采样得到的一个随机数。节点的适应性一旦被赋值，便不再改变。



在线资源5.1　比安科尼 - 巴拉巴西模型

视频中展示了一个生长过程中的网络。网络中，每个新节点被赋予一个随机的适应性参数，参数取值以节点颜色表示。每个新节点按照广义偏好连接——公式5.1 来选择与哪些节点相连。这使得节点度的增长速度正比于其适应性。图中，节点大小正比于度，随着时间推移，适应性最高的节点将成为度最大的枢纽节点。感谢王大顺提供的视频。

扫描二维码下载“湛庐阅读”App，回复书名，观看生长中的网络是怎样的。

2.偏好连接

一个新节点连接到节点i的概率正比于节点i的度k_i和适应性η_i的乘积：

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j} \qquad (5.1)$$

公式5.1 中，Π_i 和k_i 的依赖关系体现了大度节点的高可见性，这意味着新节点更可能选择和大度节点相连。Π_i 对η_i 的依赖意味着，对于度相同的两个节点，具有更高适应性的节点会有较大的概率获得新链接。因此，公式5.1 保证了，对于最初只有少数几个链接的“年轻”节点而言，只要其适应性比其他节点高，也可以快速地获得链接。

度的动态过程

可以使用连续介质理论来预测每个节点的度随时间的演化。根据公式5.1 ，节点i 的度的变化速度可以记为：

$$\frac{\partial k_i}{\partial t} = m\frac{\eta_i k_i}{\sum_j \eta_j k_j} \tag{5.2}$$

假设k_i 的时间演化服从幂指数β（η_i）依赖于其适应性的幂律分布（图5-2 ）：

$$k(t,t_i,\eta_i) = m\left(\frac{t}{t_i}\right)^{\beta(\eta_i)} \tag{5.3}$$

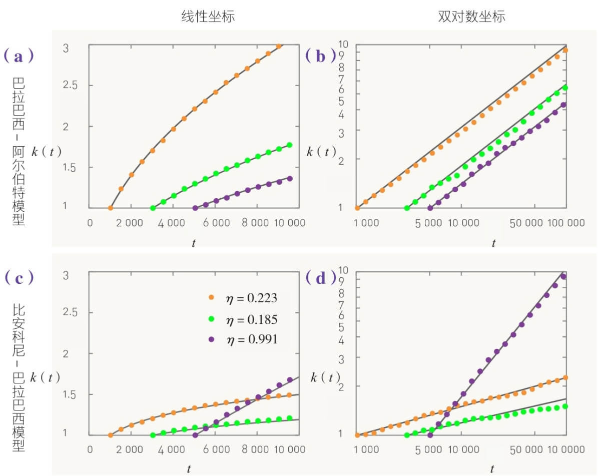


图5-2　比安科尼 - 巴拉巴西模型中的竞争

（a）　在巴拉巴西 - 阿尔伯特模型中，所有节点的度以相同的速度增长。因此，在

任何时候，早加入网络的节点都比晚加入的节点具有更大的度。图中展示了在时刻t_i =1 000, 3 000, 5 000时加入网络的节点的度随时间变化的情况。从图中可以看到，晚加入的节点不可能超过早加入的节点 [4]，[5]。

（**b**） 双对数坐标下的图a。从中可以看出，每个节点都服从幂指数为β =1/2的增长率（公式**4.7**）。

（**c**） 在比安科尼 - 巴拉巴西模型中，节点度的增长速度依赖于各自的适应性。因此较晚加入的节点，只要其适应性较高（紫色），亦有可能超过较早加入的节点。

（**d**） 双对数坐标下的图c。从中可以看出，每个节点的增长速度服从不同的幂律，幂指数依赖于各自的适应性——公式**5.3** 和公式**5.4** 。

图a、b中，每条曲线是同一适应性取值下的100次独立实验的平均值。

将公式**5.3** 代入公式**5.2** ，我们发现度的动态变化过程中的幂指数满足（进阶阅读**5.A** ）：

$$\beta(\eta)=\frac{\eta}{C} \quad (5.4)$$

其中，

$$C=\int \rho(\eta)\frac{\eta}{1-\beta(\eta)}d\eta \quad (5.5)$$

巴拉巴西 - 阿尔伯特模型中，β =1/2，因此每个节点的度按照时间的平方根速度增长。根据公式**5.4** ，比安科尼 - 巴拉巴西模型中，幂指数正比于节点的适应性η ，则节点具有各自的幂指数。因此，适应性较高的节点，度增长较快。假如时间足够长，适应性高的节点最终会将适应性低的节点甩在身后（图**5-2** ）。Facebook为该现象提供了典型案例：具有高吸引力的一位后来者，获得链接的速度要快于其竞争者，最终成为整个网络中最大的枢纽节点。

度分布

使用连续介质理论，我们可以计算出由比安科尼 - 巴拉巴西模型生成的网络的度分布（进阶阅读**5.A** ）：

$$p_k \approx C\int \frac{\rho(\eta)}{\eta}\left(\frac{m}{k}\right)^{\frac{C}{\eta}+1} d\eta \tag{5.6}$$

公式5.6 是多个幂律分布的加权和。这意味着p_k 依赖于适应性分布ρ（η）的精确形式。为了展示该模型的性质，我们使用公式5.4 和公式5.6 计算出如下两种适应性分布时的β（η）和p_k ：

1.等适应性

当所有节点的适应性相同时，比安科尼 - 巴拉巴西模型退化为巴拉巴西 - 阿尔伯特模型。实际上，这个情形对应着ρ（η）＝δ（η -1），即所有节点的适应性都为η ＝1。在这种情况下，根据公式5.5 得到C＝2，根据公式5.4 得到β ＝1/2。进而，由公式5.6 可以得到$p_k \sim k^{-3}$，即巴拉巴西 - 阿尔伯特模型中广为人知的度分布形式。

2.均匀分布的适应性

更值得关注的是，节点适应性各不相同时该模型的表现。我们假设η 均匀分布于区间[0,1]上。在该情况下，根据公式5.5 可以得到：

$$\exp(-2/C) = 1 - 1/C \tag{5.7}$$

求解公式5.7 得到C 的数值解C^* ＝1.255。因此，根据公式5.4，每个节点i 具有不同的幂指数$\beta(\eta_i) = \eta_i / C^*$ 。

使用公式5.6 ，我们得出：

$$p_k \sim \int_0^1 \frac{C^*}{\eta}\frac{1}{k^{1+C^*/\eta}} d\eta \sim \frac{k^{-(1+C^*)}}{\ln k} \tag{5.8}$$

公式5.8 表明，度分布服从度指数γ ＝2.255的幂律分布。然而，此时的度分布并非严格服从幂律分布，而是服从受对数倒数项1/ln$k$修正了的幂律分布。

图5-2 和图5-3 中的仿真结果为上述计算提供了数值上的支持。仿真结果验证了，对于每个η 均有k_i（t）服从幂律分布，且幂指数β（η

)关于η 递增。如图**5-3a** 所示，测量到的幂指数与根据公式**5.4** 计算出的幂指数精确吻合。图**5-3b** 也证实了公式**5.8** 和数值模拟结果的吻合。

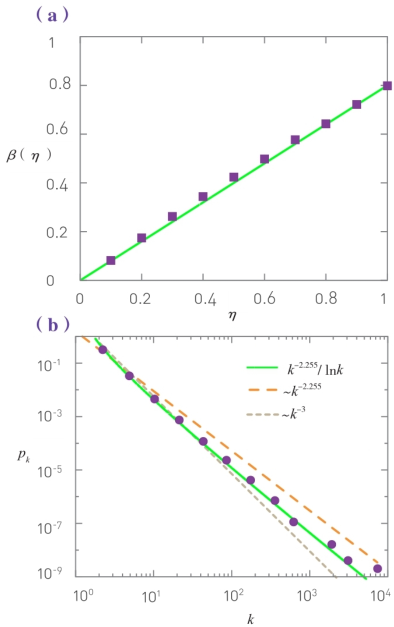


图**5-3**　刻画比安科尼 - 巴拉巴西模型

（**a**） 当η 服从均匀分布ρ (η)时，测量得到的幂指数β (η)。四边形是数值模拟的结果，实线为β (η)＝η /1.255的解析结果。

（**b**） 在m ＝2、N ＝10^6 、适应性η 服从区间[0,1]上的均匀分布时，比安科尼 - 巴拉巴西模型所生成网络的实测度分布。绿色实线对应γ ＝2.255时的公式**5.8** 。长虚线表示没有对数修正的p_k ～$k^{-2.255}$ ，短虚线表示等适应性时的度分布p_k ～k^{-3} ，即全同适应性的期望标度。注意，这三者中公式**5.8** 是最优拟合。

综上所述，比安科尼 - 巴拉巴西模型可以解释这一现象：内在属性不同的节点获取链接的速度不同。该模型预言，节点的生长速度取决于其适应性η ，我们可以由此计算出度分布对适应性分布ρ (η)的依赖关系。

5.3

测量适应性

测量节点的适应性可以帮助我们识别出即将流行起来的网站、将来具有影响力的研究论文或即将成为明星的演员（边栏**5.1**）。然而，我们判断适应性的能力是很容易出错的。例如，在判断一个关于相扑运动的网页的适应性时：一小部分人觉得相扑很迷人，大多数人表示无所谓，甚至有些人觉得相扑很奇怪。因此，不同的人对同一节点的适应性的看法可能大相径庭，而这是难以避免的。

边栏**5.1**

适应性的基因起源

适应性——在社交网络中的交友能力，是否根源于基因呢？为了回答这个问题，研究人员调查了1 110名学龄双胞胎的社交网络[6]，[7]。通过一种用于判断表征和行为的可遗传性的技术，他们发现：

- 基因因素可以解释学生入度差异中的46%（一个学生的入度是指有多少名学生视其为朋友）。
- 基因因素对学生的出度没有显著影响（一个学生的出度是指其视多少名学生为朋友）。

这表明，一个人的适应性（获得朋友的能力）是可遗传的。换言之，在社交网络中，适应性具有基因起源。另一份关于基因特征和一个人的受欢迎程度之间关系的研究也支持这一结论[8]。

根据公式**5.1**，适应性并非由某一个人给定的，而是反映了网络整体对于某个节点相对于其他节点的重要性的群体判断。因此，我们可以通过将一个节点的度随时间演化情况与网络中其他节点的度随时间演化情况进行对比，来确定该节点的适应性。本节我们将阐述，如果掌握了每个节点的度随时间演化情况，比安科尼－巴拉巴西模型的量化框架就可以使我们据此判断出每个节点的适应性。

对公式5.3　两边取对数，可以得到节点度的增长速度与节点适应性之间的关系：

$$\ln k(t,t_i,\eta_i) = \beta(\eta_i)\ln t + B_i \quad (5.9)$$

其中$B_i = \ln\left(m / t_i^{\beta(\eta_i)}\right)$是一个与时间无关的参数。因此，ln$k$（$t$，$t_i$，$\eta_i$）的斜率是关于幂指数$\beta$（$\eta_i$）的线性函数。反过来，根据公式5.4，$\beta$（$\eta_i$）线性依赖于$\eta_i$。因此，如果我们能追踪大部分节点的度随时间演化情况，幂指数β（η_i）的分布将等同于适应性分布ρ（η）。

网页的适应性

在万维网中，节点的适应性可以系统地测量出来。测量所基于的数据集，是连续13个月对2　200万个网页的链接进行每月采集而得到的[9]。虽然大部分节点（网页）的度在数据采集期间保持不变，还是有约6.5%的节点的度展现出了充分的变化，可以根据公式5.9　来确定其幂指数。得到的适应性分布ρ（η）具有指数形式（图5-4　），表明高适应性的节点是非常罕见的。

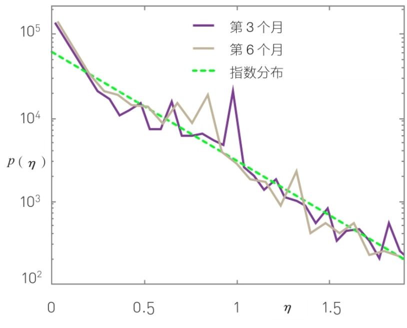


图5-4　万维网的适应性分布

根据大量网页的链接数随时间演化情况测量得到的适应性分布。测量结果表明，每个节点的度与时间成幂律依赖关系，这与公式5.3　是一致的。每条线的斜率为β（η_i），对应于节点的适应性η_i乘以一个常数——公式5.4　。图中给出了相隔三个月的两次测量结果，表明适应性分布是与时间无关的。虚线表明适应性分布可以通过指数分布进行很好的拟合。[9]

我们得到的适应性分布的形状有些出乎意料——人们直觉上会觉得不同网页的适应性应该彼此相差很大。例如，谷歌的吸引力应该远远大于我的个人主页。然而ρ（η）的指数形式表明，网页的适应性是有界的，只在一个相对较窄的范围内变化。因此，我们观测到两个网页的度相差很大，是由系统的动态演化过程决定的：适应性的微小差异会被网络生长机制和偏好连接机制放大，最终使适应性略高的节点具有非常大的度。

为了展示这一放大效应，我们考虑两个同时进入网络的节点，二者具有不同的适应性：$\eta_2>\eta_1$。根据公式**5.3**和公式**5.4**，经过较长的时间t之后，二者度的相对差异可以记为：

$$\frac{k_2-k_1}{k_1}\sim t^{\frac{\eta_2-\eta_1}{C}} \tag{5.10}$$

因此，虽然η_2和η_1的差异可能很小，但是经过很长一段时间后（t较大时），两个节点的度之间的差异可以变得非常大。

科学出版物的适应性

在一些网络中，节点度的动态演化过程要比公式**5.3**刻画的更为复杂。为了测量节点的适应性，首先需要精确刻画节点度的增长规律。接下来，我们将通过测量科学出版物的适应性来阐述其节点度的动态演化过程，从而预测科学出版物的未来影响力。

大部分科研论文很少被引用，而少数论文则被引用了成千上万次[10]。影响力的差异反映了不同论文在新颖性和重要性方面的差异。通常来说，一篇科研论文i在发表后的t时刻被引用的概率是[11]：

$$\Pi_i\sim\eta_i c_i^t P_i\left(t\right) \tag{5.11}$$

其中，论文的适应性η_i表示论文中的科学发现的新颖性和重要性；c_i^t是论文i在发表后的t时刻之前累计被引用的次数，用于刻画“被引用较多的论文比被引用较少的论文更有可能被引用”这一偏好连接现象。公式**5.11**中的最后一项刻画了论文的新颖性随时间衰减

的趋势——论文的新颖性会因后续相关论文的发表而逐渐衰减 [11] ，[12] 。实际测量表明，这一衰减过程服从对数正态的形式：

$$P_i(t)=\frac{1}{\sqrt{2\pi}t\sigma_i}e^{-\frac{(\ln t-\mu_i)^2}{2\sigma_i^2}} \tag{5.12}$$

求解公式5.11 背后的主方程，我们可得论文引用数关于时间的函数：

$$c_i^t=m\left(e^{\frac{\beta\eta_i}{A}\Phi\left(\frac{\ln t-\mu_i}{\sigma_i}\right)}-1\right) \tag{5.13}$$

其中，

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-y^2/2}dy \tag{5.14}$$

是累积正态分布，m 、β 和A 是全局参数。

从公式5.13 可以看出，论文i 的引用历史可以通过三个参数来刻画：即时性μ_i 决定一篇论文到达引用峰值所需的时间；持久性σ_i 刻画衰减速度；最重要的量是相对适应性 $\eta_i^{'}\equiv\eta_i\beta/A$ ，它度量了一篇论文相对于其他论文的重要程度，并决定了论文的最终影响力（边栏5.2 ）。

我们使用公式5.13 拟合一个期刊上每篇论文的引用记录，从而得到这个期刊的论文适应性分布（图5-5 ）。测量表明，细胞生物学顶级期刊《细胞》的适应性分布偏向右侧，表明在该期刊上发表的论文具有较高的适应性。这并不令人感到意外，因为《细胞》是所有期刊中影响因子最高的几个期刊之一。相比而言，《物理评论B》的适应性分布偏向左侧，表明在该期刊上发表的论文具有相对较低的适应性。

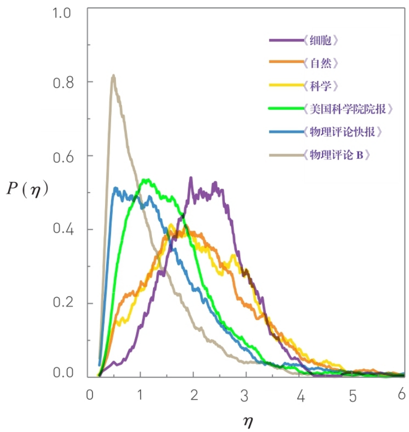


图5-5 科研论文的适应性分布

1990年6个期刊发表的所有论文的适应性分布。每篇论文的适应性根据公式5.13拟合论文发表后10年内的引用数得到。其中，有两种物理学期刊——《物理评论B》（*Physical Review B*）和《物理评论快报》，一种生物学期刊——《细胞》，三种跨领域期刊——《自然》《科学》《美国科学院院报》。

我们观测到的适应性分布相对偏移程度不同。从中可以看出，《细胞》上的论文具有最高的适应性，随后依次是《自然》《科学》《美国科学院院报》《物理评论快报》《物理评论B》。[11]

总而言之，比安科尼 - 巴拉巴西模型提供的框架使我们可以据此确定每个节点的适应性，以及适应性分布ρ（η）的形状。测量表明，适应性分布通常是指数限界的，这意味着不同节点之间的适应性相差不大。随着时间推移，适应性的差异会被放大，导致万维网中节点的指入链接数服从幂律分布，而引文网络中的论文引用数展现出较宽的分布。

边栏5.2

最终影响力

引用数仅仅反映了一篇科研论文过去的影响力。单凭引用数，我们无法知道这篇论文的影响力是已经确定，还是会持续增长。为了评估一篇论文的真实影响力，我们需要知道这篇论文在整个生命

周期内可以获得多少引用。公式**5.11** 和公式**5.14** 可以让我们预测论文的最终影响力。实际上，对公式**5.13** 取极限$t \to \infty$可得[11] ：

$$c_i^{\infty} = m\left(e^{\eta_i} - 1\right) \quad (5.15)$$

因此，尽管有很多因素会影响一篇科研论文的引用历史，但其最终影响力完全取决于其适应性η_i 。由于适应性可以使用公式**5.13** 拟合一篇论文之前的引用记录而得到，因此我们可以用公式**5.15** 来预测一篇论文的最终影响力（图**5-6** ）。

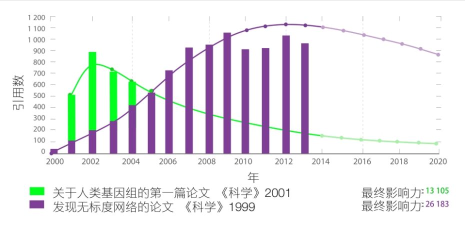


图5-6　预测最终影响力

人类基因组的第一篇论文（Venter et al [13] ）和发现无标度网络的论文（Barabási and Albert [14] ）的每年引用记录。二者的早期影响力相差很大：根据Web of Science统计，关于人类基因组的这篇论文在发表后两年内被引用1 400次，而无标度网络的论文仅仅被引用了120次。长期来看，二者的引用变化情况也很不一样：与超过85%的科研论文一样，人类基因组论文在第二年达到引用峰值；相反，在发表后大约10年内，无标度网络论文的引用逐年增长。

图中的连续曲线对应公式**5.13** 拟合引用记录的结果，我们可以据此确定论文未来的引用数和最终影响力。最终影响力即为$t \to \infty$时曲线下方的面积。根据公式**5.15** ，人类基因组论文的最终影响力为13 105，而无标度论文的最终影响力为26 183。因此，一篇论文的早期引用数并不能很好地预测其最终影响力。

5.4
玻色－爱因斯坦凝聚

在上一节中，我们发现网页的适应性分布服从简单的指数分布（图**5-4** ），而科研论文的适应性分布有一个峰值（图**5-5** ）。这些适应性分布之间的差异提出了一个重要问题：ρ (η)的形状是如何影

响网络拓扑的？

从技术上讲，将p_k 和ρ（η ）联系在一起的公式**5.6** 已经给出了答案。不过，直到我们发现有些网络存在玻色－爱因斯坦凝聚之后，适应性分布的影响才真正被意识到。在本节中，我们将讨论导致这一发现的各种“映射”，以及该发现对网络拓扑的影响[15]。

首先，我们在比安科尼－巴拉巴西模型和玻色气体之间建立起一个形式化的联系（图**5-7** ）。玻色气体的性质在物理学中已经得到了广泛研究。

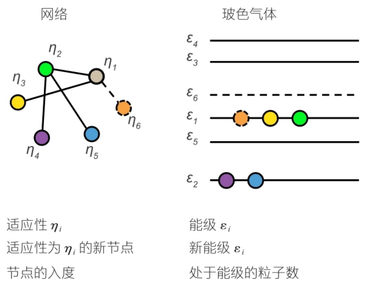


图**5-7** 将网络映射到玻色气体

网络

一个包含6个节点的网络，节点的适应性η_i 各不相同，用颜色加以区分。节点的适应性由适应性分布ρ (η)生成。

玻色气体

将每个适应性η 映射到一个能级ε ，得到一个具有随机能级的玻色气体。从一个新节点i 指向节点j 的链接对应着处于能级ε_j 的一个粒子。

增长

网络增长过程中不断有新节点加入，例如具有适应性η_6 的橙色节点。当$m=1$ 时，根据公式**5.1** ，新节点连接到灰色节点（虚线所示）。在玻色气体中，这对应着增加了一个新的能级ε_6 （虚线所示）和一个处于能级ε_1 的粒子——ε_1 为灰色节点1所在的能级。

- 适应性→能量

我们给适应性为η_i 的每个节点赋予一个能量ε_i ，赋值方式为

$$\varepsilon_i = \frac{1}{\beta_T} \log \eta_i \quad (5.16)$$

在物理系统中，β_T 表示温度的倒数。我们使用下标T 来区分β_T 和前面使用的幂指数β 。根据公式5.16 ，网络中每个节点都对应着玻色气体中的能级。节点的适应性越高，能量越低。

- 链接→粒子

对应于从节点i 指向节点j 的每条链接，我们有一个处于能级ε_j 的粒子。

- 节点→能级

一个具有m 条链接的新节点加入网络，对应于在玻色气体中增加一个新的能级ε_j 和m 个新粒子。每个新增的粒子所处的能级为新节点的相应链接指向的节点所对应的能级。

基于上述映射关系，通过数学分析可知，最终的气体中每个能级上的粒子数服从玻色统计量（边栏5.3 ）。该统计量是由萨特延德拉·纳特·玻色（Satyendra Nath Bose）于1924年推导得出的。因此，适应性模型中的链接和量子气体中的亚原子粒子的行为是一样的。

边栏5.3

从适应性到玻色气体

对于玻色气体（图5-7 ），一个粒子处于能级ε_i 的概率是

$$\Pi_i = \frac{e^{-\beta_T \varepsilon_i} k_i}{\sum_j e^{-\beta_T \varepsilon_j} k_j} \quad (5.17)$$

因此，能级ε_i 上的粒子数的增长速度为[15] ：

$$\frac{\partial k_i(\varepsilon_i, t, t_i)}{\partial t} = m \frac{e^{-\beta_T \varepsilon_i} k_i(\varepsilon_i, t, t_i)}{Z_t} \quad (5.18)$$

其中，$k_i(\varepsilon_i, t, t_i)$是能级$\varepsilon_i$上的粒子数，而

$$Z_t = \sum_{j=1}^{t} t e^{-\beta_T \varepsilon_j} k_j(\varepsilon_j, t, t_j)$$

是配分函数。方程**5.18** 的解为：

$$k_i(\varepsilon_i, t, t_i) = m\left(\frac{t}{t_i}\right)^{f(\varepsilon_i)} \quad (5.19)$$

其中，$f(\varepsilon) = e^{-\beta_T(\varepsilon-\mu)}$，而$\mu$ 是化学势，满足

$$\int \deg(\varepsilon)\frac{1}{e^{\beta_T(\varepsilon-\mu)}-1} = 1 \quad (5.20)$$

这里，deg（ε）是能级ε 的简并度。方程**5.20** 意味着在$t \to \infty$的极限条件下，能级为ε 的粒子数服从玻色统计量

$$n(\varepsilon) = \frac{1}{e^{\beta_T(\varepsilon-\mu)}-1} \quad (5.21)$$

从适应性到玻色气体的映射证明了比安科尼－巴拉巴西模型中节点的度服从玻色统计量。

从适应性到玻色气体的映射是精确的，这一映射预言了两种不同形态的存在[15]，[16]：

无标度形态

对于大多数适应性分布而言，网络表现出“富者愈富”的趋势，即每个节点的度最终由其适应性决定。虽然最大适应性的节点不可避免地将成为最大度的枢纽节点，但在这一形态中的任何时候，度分布都服从幂律，表明所生成的网络具有无标度的拓扑。因此，最大度的枢纽节点的度服从公式**3.18**，其增长是亚线性的。紧随最大度枢纽节点之后的是一些度略小的枢纽节点，其链接数与最大适应性的节点相差无几（图**5-9a**）。**5.2**节 中讨论的适应性均匀分布的模型就处于无标度形态。

玻色 - 爱因斯坦凝聚

将比安科尼 - 巴拉巴西模型映射到玻色气体，最令人意想不到的结果是：有些适应性分布会导致网络中出现玻色 - 爱因斯坦凝聚。在玻色 - 爱因斯坦凝聚中，所有粒子聚集到最低能级，而其他能级全部为空（边栏**5.4** ）。

边栏5.4

玻色 - 爱因斯坦凝聚

在经典物理中，一个运动粒子的动能$E=mv^2/2$可以为介于0（静止状态）和一个非常大的E（高速运动状态）之间的任意值，而且允许任意数量的粒子具有同样的能量E，只要这些粒子的速度v相同。在量子力学中，能量是量子化的，只能取某一些离散（量子化）的值。量子力学中有两类粒子：费米粒子（例如电子）和玻色粒子（例如光子）。同一系统中的不同费米粒子不能具有同样的能量，一个能级只能被一个电子占据（图**5-8a** ）。与之相反，可以有任意多数量的玻色粒子占据同一个能级（图**5-8b** ）。

在高温下，热扰动使得粒子具有不同的能量，此时费米气体和玻色气体的差异可以忽略（图**5-8a**，**b** ）。低温会迫使所有粒子占据尽可能低的能级，此时费米气体和玻色气体的差异变得显著起来。低温下的费米气体中，粒子从低到高逐渐填充能级，就像把水倒进花瓶中一样（图**5-8c** ）。然而，由于任意多的玻色粒子可以具有同样的能量，玻色粒子可以全部聚集在最低能级（图**5-8d** ）。换言之，无论我们往花瓶里倒入多少“玻色液体”，这些液体都会聚集在瓶底，不会填满花瓶。这一现象被称为玻色 - 爱因斯坦凝聚，由爱因斯坦在1924年首先提出。玻色 - 爱因斯坦凝聚的实验证据直到1995年才出现，并获得了2001年的诺贝尔物理学奖。

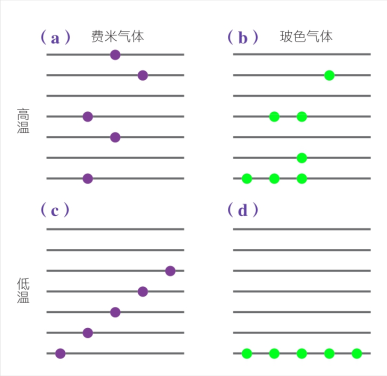


图5-8 玻色统计和费米统计

在费米气体（图a和图c）中，每个能级上只允许存在一个粒子，而在玻色气体（图b和图d）中没有这一限制。在高温下，很难分辨两种气体的差异。在低温下，每个粒子倾向于占据尽可能低的能级，两种气体的差异变得显著起来。

在网络中，玻色-爱因斯坦凝聚意味着最大适应性的节点拥有全部链接的很大一部分，会成为一个超级枢纽节点（图**5-9b**）。这样的网络不再是无标度形态，而是具有中心辐射拓扑。在这一形态中，富者愈富的机制占据主导地位，表现出赢家通吃现象，这与无标度拓扑有着本质上的不同。因此，网络将失去无标度的特性。

在物理系统中，玻色气体的温度降低到临界温度以下会发生玻色-爱因斯坦凝聚（边栏**5.4**）。在网络中，公式**5.16**中的温度β_T是一个哑变量——在包括度分布p_k在内的所有拓扑相关的物理量中不起作用。因此，玻色-爱因斯坦凝聚出现与否只取决于适应性分布ρ（η）的形式。一个网络如果要发生玻色-爱因斯坦凝聚，其适应性分布需要满足如下条件：

$$\int_{\eta_{\min}}^{\eta_{\max}} \frac{\eta\rho(\eta)}{1-\eta} d\eta < 1$$

能够产生玻色-爱因斯坦凝聚的适应性分布为：

$$\rho(\eta) = (1-\zeta)(1-\eta)^{\zeta} \qquad (5.22)$$

这里，可以通过变化ζ的值引发玻色-爱因斯坦凝聚（图**5-9**）。

实际上，公式5.20 是否有解取决于能量分布的函数形式deg (ε)，而后者是由ρ (η)的形状决定的。具体而言，对于给定的deg (ε)，如果公式5.22 没有非负解，玻色 - 爱因斯坦凝聚就会出现——粒子聚集在最低能级上。

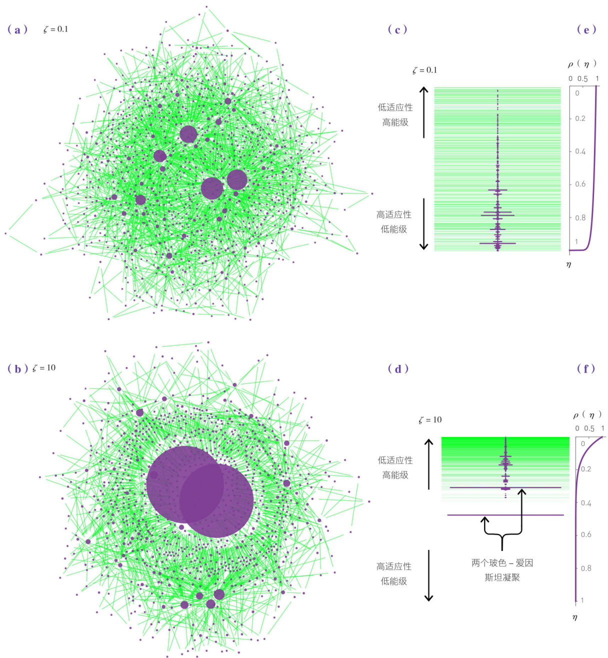


图5-9　网络中的玻色 - 爱因斯坦凝聚

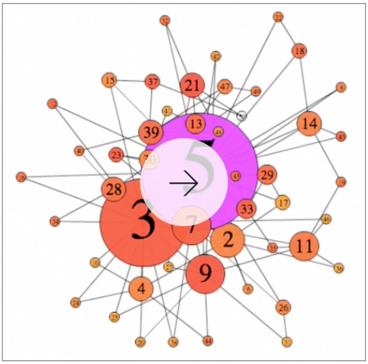
（a）（b）　无标度网络（a）和发生了玻色 - 爱因斯坦凝聚的网络（b）。两个网络都由比安科尼 - 巴拉巴西模型生成，适应性分布ρ(η) 都服从公式5.22 ，只是参数ζ不同。注意，在图b所示的玻色 - 爱因斯坦凝聚网络中，有两个大小相近的枢纽节点。

（c）（d）由比安科尼 - 巴拉巴西模型生成的$m = 2$、$N = 1\ 000$的网络所对应的能级（绿色线）和处于每个能级上的粒子数（紫色点）。能级对应图a、图b展示的网络中节点的适应性。每个粒子对应网络中的一条链接，粒子处于该链接所指向节点的能级。由于我们不允许同一对节点之间存在多条链接，导致图d中存在两个能级拥有大量粒子，对应图b中的两个枢纽节点。

（e）（f）根据公式5.22 得出的适应性分布ρ(η) ，两个ρ(η) 的形状差异是由参数ζ带来的，图e中$\zeta = 0.1$，而图f中$\zeta = 10$。

总而言之，适应性分布的确切形状决定了一个生长网络的拓扑。

一些适应性分布（例如均匀分布）产生无标度拓扑，而另一些$\rho(\eta)$会导致玻色-爱因斯坦凝聚。如果网络中发生了玻色-爱因斯坦凝聚，那么一个或少数几个节点将获得绝大多数链接。因此，导致无标度状态的富者愈富过程变成了赢家通吃现象。玻色-爱因斯坦凝聚对于网络结构的影响非常明显：它破坏了无标度网络特有的层级枢纽结构，让网络变成星型网络的中心辐射拓扑（图**5-9**）。



在线资源**5.2**　网络中的玻色-爱因斯坦凝聚

视频中展示了一个网络随时间演化的过程。其中，一个节点（紫色）的适应性比其他所有节点都高得多。这个高适应性的节点吸引了绝大多数链接，导致系统发生玻色-爱因斯坦凝聚。感谢王大顺提供的视频。

扫描二维码下载“湛庐阅读”App，回复书名，了解玻色-爱因斯坦凝聚是如何形成的。

5.5
演化网络

巴拉巴西-阿尔伯特模型是一个极简的模型，旨在刻画导致无标度特性出现的机制。因此，它有几个为人熟知的局限性（参见**4.11**节

）：

（**1**） 该模型断言度指数$\gamma=3$，而实验观测到的度指数取值范围为2到5之间（表**3-1** ）。

（**2**） 该模型断言度分布为幂律分布，而在真实系统中，我们观测到的度分布和严格的幂律分布之间存在系统性偏离，例如低度饱和和高度截断（边栏**3.8** ）。

（**3**） 该模型忽略了许多真实网络中明显存在的基本过程，例如添加内部链接、删除节点和链接。

上述局限性引发了许多研究，去尝试理解众多影响网络拓扑的基本过程在网络演化中所发挥的作用。在本节中，我们将系统性地扩展巴拉巴西－阿尔伯特模型，并介绍一组演化网络模型，以解释真实网络拓扑中出现的各种现象。

初始吸引力

在巴拉巴西－阿尔伯特模型中，一个孤立节点无法获得链接。根据公式**3.1** 所示的偏好连接机制，一个新节点连接到度$k=0$的节点的概率为零。而在真实网络中，即使是孤立节点也可能获得链接。实际上，每一篇新的科研论文都有一定的概率被首次引用，搬到新城市的人很快会交到新朋友。为了允许孤立节点获得链接，我们给公式**3.1** 所示的偏好连接加上一个常数：

$$\Pi(k)\sim A+k \qquad (5.23)$$

这里，常数A 被称为初始吸引力。由于$\Pi(0)\sim A$ ，初始吸引力正比于节点在下一时刻获得首个链接的概率。

对$\Pi(k)$的直接测量证实了真实网络存在初始吸引力（图**5-10** ）。当初始吸引力存在时，会出现两种结果：

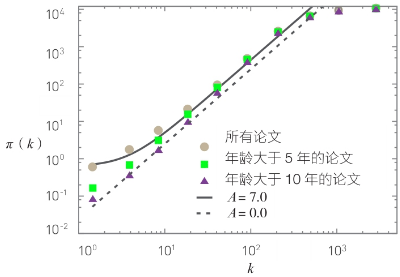


图5-10　初始吸引力

引文网络的累积偏好连接函数，即公式**4.21** 。图中展示了2007年至2008年间发表的科研论文的引用模式。曲线$\pi(k)$是使用第**4.7**节 中所描述的方法测量得到的。实线对应初始吸引力$A \sim 7.0$的情形，短虚线对应$A = 0$的情形，即没有初始吸引力的情形。$A = 7$意味着，对一篇新论文而言，计算其首次被引用的概率时相当于该论文已经有了7次引用。[19]

（1）增大度指数

如果我们把巴拉巴西 - 阿尔伯特模型中的公式**3.1** 替换为公式**5.23** ，度指数则变成[17]，[18]：

$$\gamma = 3 + \frac{A}{m} \tag{5.24}$$

因此，初始吸引力增大了度指数γ ，使网络更加同质化，且枢纽节点变小。实际上，初始吸引力是给连接到一个节点的概率增加了一个随机项。这个随机项对低度节点的概率提升更为明显，因而削弱了偏好连接的作用。对于高度节点，公式**5.23** 中的初始吸引力A 的作用可以忽略。

（2）导致低度饱和

求解连续介质方程表明，按照公式**5.23** 生成的网络，其度分布并不服从严格的幂律，而是具有如下形式：

$$p_k = C\left(k + A\right)^{-\gamma} \tag{5.25}$$

因此，初始吸引力导致在$k < A$ 时出现低度饱和，起到了公式

3.39 中k_{sat}的作用。造成这一低度饱和现象的原因是：初始吸引力提高了新节点连接到低度节点的概率，从而提升了低度节点的度。高度节点（$k \gg A$）的度分布仍然服从幂律，因为初始吸引力对这些节点的连接概率改变很小。

内部链接

在很多网络中，新链接并不总是由新节点带来的，网络中的已有节点之间也可以形成链接——被称为内部链接。例如，万维网中绝大多数新的链接是内部链接，即已有网页上新增的超链接。与之类似，社会网络中几乎所有的新增朋友关系都属于内部链接。

测量表明，在科学合作网络中，内部链接遵循双重偏好连接：度为k的节点和度为k'的节点之间新增一条内部链接的概率为[20]：

$$\Pi(k,k') \sim (A+Bk)(A+Bk') \quad (5.26)$$

为了理解内部链接带来的影响，我们将考察公式5.26的两个极限情况：

（1）双重偏好连接（$A=0$）

考虑巴拉巴西－阿尔伯特模型的一个扩展：在每个时间步，我们增加一个具有m条链接的新节点，然后根据公式5.26增加n条内部链接，这里假设$A=0$。因此，出现一条新链接的概率正比于其连接的两个节点的度。最终生成的网络，其度指数为[21]，[22]：

$$\gamma = 2 + \frac{m}{m+2n} \quad (5.27)$$

可以看出，度指数γ介于2与3之间。这意味着，双重偏好连接将度指数从3降到了2，从而增加了网络的异质性。实际上，内部链接倾向于在枢纽节点之间形成，从而让枢纽节点变得更大，进一步削弱低度节点。

（2）随机内部连接（$B=0$）

在这种情况下，内部链接无视节点的度。因此，内部链接在随机选择的节点对间形成。在巴拉巴西 - 阿尔伯特模型中，假如我们在每次新增一个节点后随机挑选n 个节点对用于增加内部链接，则最终得到的网络的度指数为[22] ：

$$\gamma = 3 + \frac{2n}{m} \tag{5.28}$$

因此，对任意n 和m 的组合，我们有$\gamma \geq 3$。这意味着，最终得到的网络比没有内部链接的网络更加同质化。实际上，随机增加内部链接模拟的是随机网络的过程，因此使节点的度之间的差异变小。

节点删除

许多真实系统中存在节点和链接消失的情况。例如，员工离职后该节点即从组织结构网络中删除，网页移除后该节点即从万维网中删除。同时，也有一些网络几乎不可能移除节点（图**5-11**）。

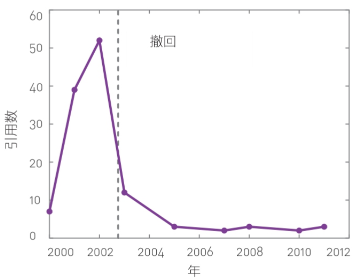


图**5-11**　无法删除的节点

简·亨德里克·舍恩（Jan Hendrik Schön）发表在《科学》杂志上的一篇论文[23] 的引用记录，展示了从引文网络中移除一个节点有多困难。在发表了一系列半导体突破性进展后，舍恩成了这一领域的新星。他的发表速度非常惊人：在2001年，他每8天就发表一篇合作科研论文，而这些论文大多发表在《科学》和《自然》这样的顶级期刊上。

在舍恩发表了一篇关于单分子半导体的突破性发现的论文后，研究人员发现他记录的两个实验有不同的温度和完全一致的噪声[24] 。随后的质疑使得舍恩工作的贝尔实验室的母公司朗讯科技对他展开了正式调查。最终舍恩承认伪造数据。他的数十篇论文被撤回，包括图中这篇展示引用模式的论文。

虽然论文的撤回导致引用急剧减少，这些论文在被官方“删除”后仍然继续被引用，如上图所示。这表明从引文网络中移除一个节点实际上是不可能的。

我们从巴拉巴西 - 阿尔伯特模型出发，探索节点删除的影响。在每个时间步，我们增加一个具有m 条链接的新节点，并且以速率r 删除一个节点。根据r 的取值不同，我们观察到三种不同的形态[25-30]：

（**1**）无标度形态

当$r<1$时，删除的节点数少于新增的节点数，因此网络持续增长。在这种情况下，网络是无标度的，其度指数为：

$$\gamma=3+\frac{2r}{1-r} \quad (5.29)$$

因此，随机删除节点会增大γ，使网络趋于同质化。

（**2**）指数形态

当$r=1$时，新增和删除节点的速度相等，因此网络将维持固定大小（N 为常数）。在这种情况下，网络的无标度性质会消失。实际上，当$r\rightarrow 1$时，公式**5.29** 中的γ 会趋于无穷大。

（**3**）衰退网络

当$r>1$时，删除节点的数量多于新增节点的数量，网络规模逐渐衰退（边栏**5.5**）。衰退网络在一些场合下会出现。例如，阿尔茨海默病的研究重点是，随着年龄增长神经元会逐渐消失，而生态学也在探讨栖息地的逐渐缩减[31-33]。衰退网络的一个典型例子是电报网络。电报曾是19世纪下半叶和20世纪早期的主要长途通信手段。它曾是一个增长网络：美国的电报线从1846年的64千米增长到1852年的37 015千米。然而，在第二次世界大战后，电报逐渐消失了。

如果节点删除还伴随着其他基本过程，网络的行为会变得相当复杂。如图**5-12** 所示，初始吸引力和节点删除同时存在的时候，网络会在无标度网络和指数网络之间发生相变。最后，需要注意的是，节点删除并不总是随机的，有可能与被删除节点的度相关（边栏**5.5**

）。

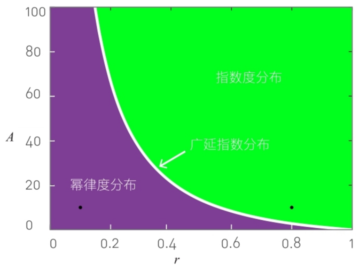


图5-12　节点删除导致的相变

节点删除与其他基本过程同时存在可能会导致有趣的拓扑形态变化。我们通过一个简单的模型来加以说明。在这个模型中，网络按照公式**5.23**　生长，同时以速率r　删除节点 [30] 。网络表现出三种不同的形态（如图所示）。其中横轴和纵轴分别表示节点移除速度r 和初始吸引力A ：

次临界节点删除：$r<r^*(A)$

当节点删除的速度低于图中白线所示的临界值$r^*(A)$时，网络是无标度的。

临界节点删除：$r=r^*(A)$

当r 取临界值$r^*(A)$时，度分布变成了广延指数分布（进阶阅读**3.A**　）。

指数网络：$r>r^*(A)$

网络丧失了无标度性质，形成指数形式的度分布。

因此，多种基本过程同时存在于网络中可能导致网络拓扑的突变。具体而言，节点删除的速度增长将导致从无标度网络到指数网络的相变。

边栏5.5

衰退的时装网络

纽约的服装业可作为衰退网络的典型例子（图**5-1**　）。其节点为设计师和承包商，他们通过每年合作制衣的关系彼此相连。随着行业的衰退，这个网络在持续缩小：网络中的最大连通分支从1985年的3　249个节点缩减到2003年的190个节点。有趣的是，网络的度分布在这段时间里保持不变。分析该网络的演化可知，衰退网络具有以下两个特性 [25] ：

- 偏好连接

在整个网络缩小的同时仍然有新节点加入这个网络。观测表明，这些新节点的连接概率服从$\Pi(k)\sim k^{\alpha}$，其中$\alpha=1.20\pm0.06$（图**5-13a**），为超线性偏好连接提供了证据（第**4.8**节）。

- 链接删除

一家公司失去一条链接的概率服从$k(t)^{-\eta}$，其中$\eta=0.41\pm0.04$。也就是说，公司失去链接的概率随着度的增加而降低（图**5-13b**）。这反映了“弱者愈弱”的现象，即拥有较少链接的公司更有可能失去链接。

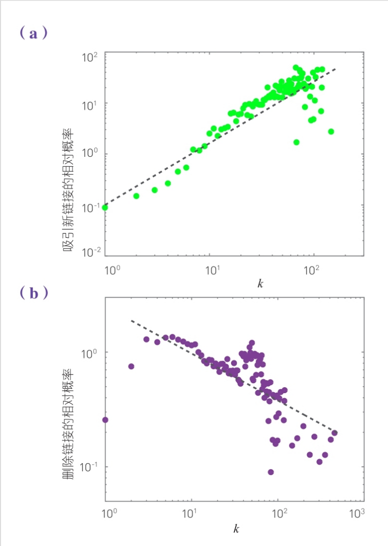


图**5-13**　时装业的衰退

（**a**）偏好连接。相对于随机增加一条链接，在t时刻成立的一家新公司连接到一个已有k条链接的公司的概率。虚线的斜率为$\alpha=1.2$。若新增链接是完全随机的，斜率大约为1。

（**b**）链接删除。相对于随机删除链接，从一个度为k的节点上删除一条链接的概率。虚线的斜率为$\eta=0.41$。若链接删除是完全随机的，对于任意k，该相对概率约为1。

总之，大多数网络存在节点消失的可能。但只要网络持续增长，其无标度特性就可以保留。然而，度指数的取值取决于节点删除过程的具体情况。

加速增长

在此前所讨论的模型中，链接数随着节点数的增长而呈线性增长。换言之，我们假设$L=\langle k\rangle N/2$，其中$\langle k\rangle$与时间无关。对于大多数真实网络来说，这是一个合理的假设。然而，在某些真实网络中，链接数比节点数N增长得快。这一现象被称为加速增长。例如，互联网的平均度从1997年11月的$\langle k\rangle=3.42$增长到1998年12月的3.96 [34]；万维网的平均度在5个月里从7.22增长到7.86 [35]，[36]；代谢网络中代谢物的平均度几乎随着代谢物的数量呈线性增长 [37]。为了探索加速增长带来的影响，我们假设在增长网络中，每个新节点带来的链接数服从 [38-41]：

$$m(t)=m_0t^{\theta} \tag{5.30}$$

当$\theta=0$时，每个新节点具有同样的链接数；而当$\theta>0$时，网络加速增长。

将公式5.30中的加速增长引入巴拉巴西－阿尔伯特模型中，所得网络的度指数为：

$$\gamma=3+\frac{2\theta}{1-\theta} \tag{5.31}$$

因此，加速增长使度指数超过$\gamma=3$，网络变得更同质化。当$\theta=1$时，度指数发散，出现超加速增长 [39]，此时，$\langle k\rangle$随时间呈线性增长，网络不再具有无标度特性。

老化

在很多真实系统里，节点的寿命是有限的。例如，演员的职业生涯长度是有限的。同样，科学家的职业生涯——连续发表科研论文的

时期，其长度也是有限的。在这些网络中，节点并不会忽然消失，而是在一个缓慢的老化过程中慢慢消失。在这个过程中，节点吸引新链接的速度会逐渐降低[42-45]。容量限制可以导致类似的现象：节点处理链接的能力是有限的，当达到容量限制后，节点便停止接受新链接[43]。

为了理解老化的影响，我们假设一个新节点连接到节点i的概率是$\Pi(k_i, t-t_i)$，其中t_i是节点i加入网络的时间，$t-t_i$是该节点的年龄。老化效应可以建模为[42]：

$$\Pi\left(k_i, t-t_i\right) k_i\left(t-t_i\right)^{-\nu} \qquad (5.32)$$

其中，ν是一个可调的参数，用于控制连接概率和节点年龄之间的依赖关系。根据ν的取值不同，有如下三种情况：

（1）ν 为负值

如果$\nu<0$，新节点会倾向于连接到年龄大的节点。因此，负的ν强化了偏好连接的作用。在$\nu\to-\infty$的极端情况下，每个新节点都选择和最老的节点相连，形成中心辐射拓扑（图**5-14a**）。计算表明，ν为负值时仍然存在无标度性质，但度指数降到了3以下（图**5-14e**）。因此，$\nu<0$让网络更加异质化。

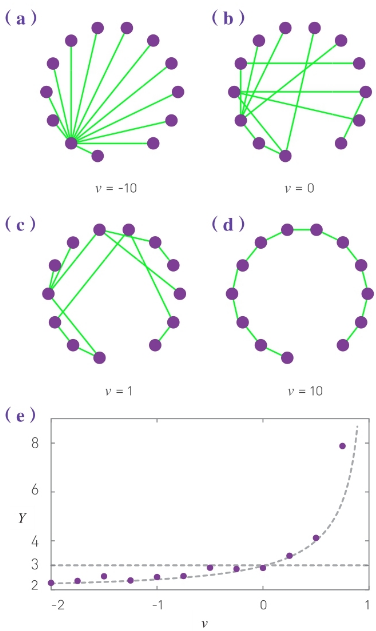


图5-14　老化的影响

（a）～（d）公式5.32 中老化指数ν不同的情况下，期望网络拓扑的示意图。在生长网络中，我们假设连接到节点的概率正比于$k\tau^{-\nu}$，其中τ 是节点的年龄。若ν 为负值，节点倾向于连接到最老的节点，使得网络变为中心辐射拓扑。若ν 为正值，最新的节点最有吸引力。当ν 非常大时，由于最后加入（最年轻）的节点总是最有吸引力的，网络会变成链式结构。简单起见，图中展示的网络假设$m=1$，实际上度指数也和m 无关。

（e）老化模型中，度指数γ 与老化指数ν 的关系。紫色标记对应的是数值模拟的结果，每个记号对应一个$N=10\ 000$且$m=1$的网络。此图是根据《物理评论快报》曾发表的一篇论文[42]重绘的。

（2）ν 为正值

此时，新节点倾向于连接到年轻的节点。在$\nu \to \infty$的极端情况下，每个新节点都连接到刚加入网络的节点（图5-14d）。即使当ν的取值不太大时，我们也可以观察到衰老的影响：当ν 逐渐接近1时，度指数发散（图5-14e）。因此，老化效应让年老的枢纽节点不再有吸引力，从而让网络趋于同质化。

（3）$\nu>1$

此时，老化效应消除了偏好连接的作用，使得网络不再具有无标度性质（图**5-14d**）。

总之，本节讨论的实验结果表明，很多基本过程可以影响一个生长网络的结构和演化过程（图**5-15**）。这些结果凸显了网络演化模型的力量：它允许我们使用自洽的数学预测框架来考察各种过程对于网络拓扑和演化的影响。

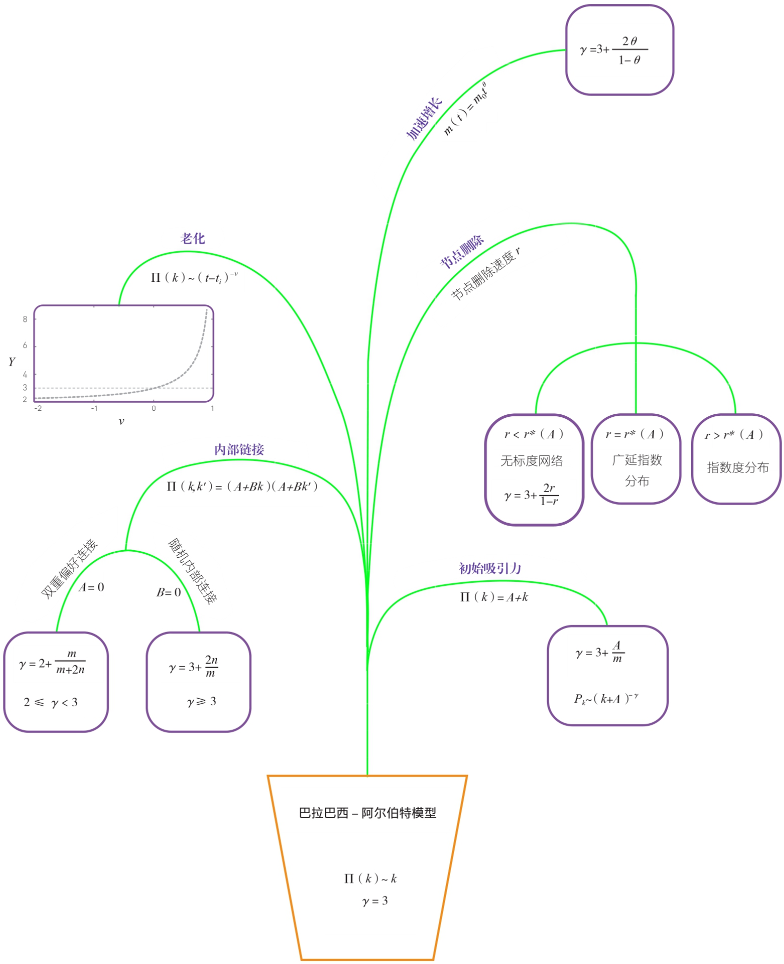


图5-15　影响网络拓扑的基本过程

本节中讨论的基本过程及其对度分布的影响。这里的每个模型都可以视为巴拉巴西－阿尔伯特模型的扩展。

5.6 小结

如本章所示，从适应性到内部链接和老化效应等许多基本过程可以影响真实网络的结构。从中我们可以学习到如何利用演化网络理论来探讨各种基本过程对网络拓扑和演化的影响。本章讨论的例子使我们得出一个关键结论：想要理解网络的结构，首先需要掌握其动态演化过程。结构是演化的结果。

本章给出的一些数学工具使我们可以更清晰地理解前面章节中遇到的诸多问题，包括如何正确拟合度分布和不同网络建模框架的作用等。接下来，我们将简要讨论其中的一些问题。

拓扑多样性

在第**3**章　中，我们讨论了尝试用纯幂律分布拟合真实网络度分布时所遇到的困难。该问题的根源在本章会变得清晰起来：如果我们理解了在网络演化中发挥作用的那些真实动态过程，便能理解真实网络度分布和纯幂律分布之间的系统性偏离。实际上，我们在本章中给出了度分布的一些解析形式：

（**1**）幂律

当网络演化完全由线性偏好连接机制决定时，根据巴拉巴西－阿尔伯特模型可知，节点的度严格服从幂律分布。而真实系统中很难观察到这样的纯幂律度分布。不过，该理想化的模型为我们理解真实网络的度分布提供了一个很好的起点。

（**2**）广延指数分布

如果偏好连接是亚线性的，则度分布服从广延指数分布（参见**4.8**节　）。在节点删除机制的临界状态，也会出现类似的度分布（图**5-12**　）。

（3）适应性分布的修正

考虑适应性时，度分布p_k 的精确形式取决于适应性分布ρ（η ），参见公式**5.6** 。例如，如公式**5.8** 所示，均匀分布的适应性导致p_k 存在一个对数修正。其他形式的ρ (η)可能产生怪异形式的p_k 。

（4）低度饱和

初始吸引力给偏好连接增加了一个随机项。从而，度分布出现了公式**5.24** 所示的低度饱和。

（5）高度截断

许多真实网络中存在的节点删除和链接删除，可能导致度分布出现指数形式的高度截断。而且，随机节点删除会大量消耗掉低度节点，从而使p_k 出现峰值。

本章所讨论的基本过程在大多数真实网络中同时存在。例如，在科学合作网络中，初始吸引力和亚线性偏好连接共同作用，还存在内部链接。由于科研人员的创造力不同，适应性也会发挥作用，一个准确的模型需要我们知道合适的适应性分布。因此，度分布会表现出低度饱和（存在初始吸引力所致）、广延指数形式的高度截断（亚线性偏好连接所致），以及由适应性分布ρ （η)的特定形式带来的一些未知修正。

一般而言，如果我们希望精确拟合度分布，首先需要构建一个能够产生p_k 函数形式的生成式模型。然而，在许多系统中，为p_k 构建一个准确的理论模型是不现实的。实际上，只要知道p_k 是具有指数限界还是具有重尾分布，通常就足够了（参见**3.9**节 ）——系统的性质主要是由这一点决定的。

模型多样性

本章给出的结果也使我们理清了前面几章中遇到的各种网络模型在网络建模中的不同作用。我们可以把这些模型分成三大类，见表

5-1 。表中总结了网络科学的三类主要模型框架，及其各自的特点：

表5-1　网络科学中的模型分类

类型	模型	特性
静态模型	埃尔德什－雷尼模型 瓦茨－斯托加茨模型	· 固定节点数 N · p_k 指数有界 · 与时间无关的静态拓扑
生成式模型	配置模型 隐参数模型	· 任意预定义的 p_k · 与时间无关的静态拓扑
演化网络模型	巴拉巴西－阿尔伯特模型 比安科尼－巴拉巴西模型 初始吸引力模型 含内部链接的模型 节点删除模型 加速增长模型 老化模型	· p_k 取决于参与网络演化的过程 · 随时间变化的网络拓扑

（**1**）静态模型

埃尔德什和雷尼的随机网络模型（第**2**章　）及瓦茨和斯托加茨的小世界网络模型（边栏**2.8**　）具有固定数量的节点，我们称其为静态模型。这两个模型都假设网络建模者的作用只是选择某种随机算法，在节点之间放置链接。理解这些网络模型的性质需要借助于埃尔德什和雷尼建立的组合图论。这两个模型所产生的网络具有有界的度分布。

（**2**）生成式模型

第**3.8**节　讨论的配置模型和隐参数模型使用预定义的度分布来生成网络。因此，这些模型并非在刻画机制，也就是说，它们并不会告诉我们为什么网络会出现某种特定的度分布。相反，这些模型只是帮助我们理解度分布是如何决定了聚团和路径长度等各种网络属性的。

（**3**）演化网络模型

这类模型刻画了网络随时间演化的机制。人们研究得最多的是巴

拉巴西－阿尔伯特模型。本章讨论的其他扩展模型同样也提供了新的见解，包括比安科尼－巴拉巴西模型和涉及内部链接、老化、节点删除、链接删除、加速增长的模型。促使人们提出这些模型的动力来自这样一种观测：如果我们能够正确地刻画对网络演化有作用的所有微观过程，就能正确地建模网络的拓扑特性。为了探索这些模型所生成网络的性质，需要使用动态方法，例如连续介质理论和速率方程方法。

每种建模方式在网络理论中都发挥着重要作用。埃尔德什－雷尼模型帮助我们检验某种网络性质是否可以通过纯随机连接来解释。如果我们只关注网络环境对于某些现象的作用，例如传播过程或网络鲁棒性，那么生成式模型是一个非常好的出发点。如果我们希望理解某种网络性质的成因，就需要采取演化网络模型从头刻画网络的形成过程了。

5.7 课后习题

习题1　加速增长

计算包含加速增长机制的、有向的巴拉巴西－阿尔伯特模型所生成网络的度指数。假设新加入节点的度随时间增长的趋势为$m(t)=t^{\theta}$。

习题2　舞会的演化网络模型

在一个不考虑性别的舞会中，每个新来的人从其他人中邀请一个人跳舞。在选择舞伴时，吸引力起决定作用：越有吸引力的人越容易收到新来者的邀请。舞会按照以下规则演化：

- 每个人对应一个节点i，被赋予一个与时间无关的吸引力η_i。
- 在每个时间步，一个新节点加入舞会。

- 新节点邀请一个已有节点跳舞，与其建立一条链接。
- 新节点选择舞伴的概率正比于舞伴的吸引力。如果舞会中已经有t 个节点了，那么节点i 收到跳舞邀请的概率为：

$$\Pi_i = \frac{\eta_i}{\sum_j \eta_j} = \frac{\eta_i}{t\langle\eta\rangle}$$

其中$\langle\eta\rangle$是平均吸引力。

（1） 推导出节点度随时间的演化情况，即一个人跳了多少支舞。

（2）推导出吸引力为η 的节点的度分布。

（3） 如果一半节点的吸引力为η ＝2，另一半节点的吸引力为η ＝1，那么经过足够长的时间后，网络的度分布是什么样的？

习题3 比安科尼－巴拉巴西模型

考虑只有两个适应性取值的比安科尼－巴拉巴西模型，即η ＝a和η ＝1。具体而言，我们假设适应性服从双德尔塔分布（double delta distribution）：

$$\rho(\eta) = \frac{1}{2}\delta(\eta - a) + \frac{1}{2}\delta(\eta - 1)$$

其中0≤a ≤1。

（1）计算度指数及其与参数a 的关系。

（2）计算网络的平稳度分布。

习题4 加性适应性

假设支配网络生长过程的偏好连接机制包含加性适应性

$$\Pi(k_i) \sim \eta_i + k_i$$

其中，每个节点有各自的适应性η_i，η_i 从适应性分布ρ（η）中采样得到。计算并讨论所得到的网络的度分布。

5.8 进阶阅读5.A 比安科尼 - 巴拉巴西模型的解析解

这一节，我们的目标是推导出比安科尼 - 巴拉巴西模型的度分布 [2]，[15]，[16]，[17]。考虑所有可能的适应性η，我们首先计算

$$\left\langle \sum_j \eta_j k_j \right\rangle$$

由于每个节点加入网络的时间t_0不同，我们可以把关于j 的求和式写成关于t_0的积分

$$\left\langle \sum_j \eta_j k_j \right\rangle = \int \rho(\eta)\eta d\eta \int_1^t k_\eta(t,t_0)dt_0 \qquad (5.34)$$

将$k_\eta(t, t_0)$替换为公式5.3，并对t_0求积分可得：

$$\left\langle \sum_j \eta_j k_j \right\rangle = \int \rho(\eta)\eta d\eta m \frac{t - t^{\beta(\eta)}}{1-\beta(\eta)} \qquad (5.35)$$

由于节点的度只能随时间增长（$\beta(\eta)>0$），且$k_i(t)$不会比t 增长得快（$\beta(\eta)<1$），因此$\beta(\eta)$是有界的，即$0<\beta(\eta)<1$。在$t\to\infty$的极限下，公式5.35 中的$t^{\beta(\eta)}$与t 相比可以忽略不计，因此有：

$$\left\langle \sum_j \eta_j k_j \right\rangle \overset{t\to\infty}{=} Cmt\left(1-O\left(t^{-\varepsilon}\right)\right) \qquad (5.36)$$

其中，$\varepsilon=(1-\max_\eta \beta(\eta))>0$且

$$C = \int \rho(\eta)\frac{\eta}{1-\beta(\eta)}d\eta \qquad (5.37)$$

记$k_\eta = k_\eta(t, t_0, \eta)$，根据公式5.36，动力学方程5.2可以写为：

$$\frac{\partial k_\eta}{\partial t} = \frac{\eta k_\eta}{Ct} \quad (5.38)$$

其解具有公式5.3的形式。另外，

$$\beta(\eta) = \frac{\eta}{C} \quad (5.39)$$

证实了公式5.3的假设是自洽的。

要完成计算，我们还需要根据公式5.37确定C的值。将$\beta(\eta)$替换为η/C，可得

$$1 = \int_0^{\eta_{\max}} \rho(\eta) \frac{1}{\frac{\eta}{C} - 1} d\eta \quad (5.40)$$

其中$\eta_{\max}$是系统中最大可能的适应性。积分5.40是奇异的。然而，由于对于任意η，存在$\beta(\eta) = \eta/C < 1$，我们有$C > \eta_{\max}$。因此，公式5.40的积分极限不会达到奇异点。注意，由于

$$Cmt = \sum_j \eta_j k_j \leq \eta_{\max} \sum_j k_j = 2mt\eta_{\max} \quad (5.41)$$

我们有$C \leq 2\eta_{\max}$。

如果存在单一的幂指数β，度分布服从度指数为$\gamma = 1/\beta + 1$的幂律分布$p_k \sim k^{-\gamma}$。然而，在比安科尼 - 巴拉巴西模型中，幂指数$\beta(\eta)$是一个分布，因此p_k是多个幂律分布的加权和。

要确定N很大时的度分布，我们首先要计算适应性为η且度大于k的节点个数，即满足$k_\eta(t) > k$的节点数。根据公式5.3，我们发现这意味着：

$$t_0 < t\left(\frac{m}{k}\right)^{C/\eta} \quad (5.42)$$

在每个时刻，恰好有一个节点加入，且每个节点以概率 $t\left(\frac{m}{k}\right)^{C/\eta}\rho(\eta)d\eta$ ρ（η）dη 具有适应性η 。因此，有个节点满足条件公式5.42 。为了得到累积分布函数（一个随机节点i 的度小于或等于k 的概率），我们有：

$$P(k)=P(k_i\le k)=1-P(k_i>k)\approx 1-\frac{\int_0^{\eta_{\max}} t\left(\frac{m}{k}\right)^{C/\eta}\rho(\eta)d\eta}{m_0+t}$$

$$\approx 1-\int_0^{\eta_{\max}}\left(\frac{m}{k}\right)^{C/\eta}\rho(\eta)d\eta \qquad (5.43)$$

其中最后一个等式在t 很大时近似成立。度分布的概率密度函数为：

$$p(k)=P'(k)=\int_0^{\eta_{\max}}\frac{C}{\eta}m^{C/\eta}k^{-(C/\eta+1)}\rho(\eta)d\eta \qquad (5.44)$$

从而得到了公式5.6 。

5.9 参考文献

[1] A.L. Barabási. *Linked: The New Science of Networks* .Perseus, Boston, 2001.

[2] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters* , 54: 436-442, 2001.

[3] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web. *Science* , 287: 2115, 2000.

[4] P.L. Krapivsky and S. Redner. Statistics of changes in lead node in connectivity-driven networks. *Phys. Rev. Lett* ., 89:258703, 2002.

[5] C. Godreche and J. M. Luck. On leaders and condensates in a growing network. *J. Stat. Mech* ., P07031, 2010.

[6] J. H. Fowler, C. T. Dawes, and N. A. Christakis. Model of Genetic Variation in Human Social Networks. *PNAS* , 106: 1720-1724, 2009.

[7] M. O. Jackson. Genetic influences on social network characteristics. *PNAS* , 106:1687-1688, 2009.

[8] S.A. Burt. Genes and popularity: Evidence of an evocative gene environment correlation. *Psychol. Sci* ., 19:112-113, 2008.

[9] J. S. Kong, N. Sarshar, and V. P. Roychowdhury. Experience versus talent shapes the structure of the Web. *PNAS* , 105:13724-9, 2008.

[10] A.-L. Barabási, C. Song, and D. Wang. Handful of papers dominates citation. *Nature* , 491:40, 2012.

[11] D. Wang, C. Song, and A.-L. Barabási. Quantifying Long term scientific impact. *Science* , 342:127-131, 2013.

[12] M. Medo, G. Cimini, and S. Gualdi. Temporal effects in the growth of networks. *Phys. Rev. Lett* ., 107:238701, 2011.

[13] C. Venter et al. The sequence of the human genome. *Science* , 291:1304-1351, 2001.

[14] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* , 286:509-512, 1999.

[15] G. Bianconi and A.-L. Barabási. Bose-Einstein condensation in complex networks. *Phys. Rev. Lett* ., 86: 5632-5635, 2001.

[16] C. Borgs, J. Chayes, C. Daskalakis, and S. Roch. First to market is not everything: analysis of preferential attachment with fitness. STOC'07, San Diego, California, 2007.

[17] S. N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett* ., 85: 4633, 2000.

[18] C. Godreche, H. Grandclaude, and J.M. Luck. Finite-time fluctuations in the degree statistics of growing networks. *J. of Stat. Phys* ., 137:1117-1146, 2009.

[19] Y.-H. Eom and S. Fortunato. Characterizing and Modeling Citation Dynamics. *PLOS ONE* , 6: e24926, 2011.

[20] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A* , 311: 590-614, 2002.

[21] R. Albert, and A.-L. Barabási. Topology of evolving networks: local events and universality. *Phys. Rev. Lett* ., 85:5234-5237, 2000.

[22] G. Goshal, L. Chi, and A.-L Barabási. Uncovering the role of elementary processes in network evolution. *Scientific Reports* , 3:1-8, 2013.

[23] J.H. Schön, Ch. Kloc, R.C. Haddon, and B. Batlogg. A superconducting field-effect switch. *Science* , 288: 656-8. 2000.

[24] D. Agin. *Junk Science: An Overdue Indictment of Government, Industry, and Faith Groups That Twist Science for Their Own Gain.* Macmillan, New York, 2007.

[25] S. Saavedra, F. Reed-Tsochas, and B. Uzzi. Asymmetric disassembly and robustness in declining networks. *PNAS* , 105:16466-16471, 2008.

[26] F. Chung and L. Lu. Coupling on-line and off-line analyses for random power-law graphs. *Int. Math* ., 1: 409-461, 2004.

[27] C. Cooper, A. Frieze, and J. Vera. Random deletion in a scale-free random graph process. *Int. Math* . 1, 463-483, 2004.

[28] S. N. Dorogovtsev and J. Mendes. Scaling behavior of developing and decaying networks. *Europhys. Lett* ., 52: 33-39, 2000.

[29] C. Moore, G. Ghoshal, and M. E. J. Newman. Exact solutions for models of evolving networks with addition and deletion of nodes. *Phys. Rev. E* , 74: 036121, 2006.

[30] H. Bauke, C. Moore, J. Rouquier, and D. Sherrington. Topological phase transition in a network model with preferential attachment and node removal. *The European Physical Journal B* , 83: 519-524, 2011.

[31] M. Pascual and J. Dunne, (eds). *Ecological Networks:Linking Structure to Dynamics in Food Webs* . Oxford University Press, Oxford, 2005.

[32] R. Sole and J. Bascompte. *Self-Organization in Complex Ecosystems* . Princeton University Press, Princeton, 2006.

[33] U. T. Srinivasan, J. A. Dunne, J. Harte, and N. D.Martinez. Response of complex food webs to realistic extinction sequencesm. *Ecology* , 88:671-682, 2007.

[34] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review* , 29: 251-262, 1999.

[35] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S.Rajagopalan, R. Stata, and A. Tomkins. Graph structure in the web. *Computer Networks* , 33: 309-320, 2000.

[36] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters. ACM TKDD07,*ACM Transactions on Knowledge Discovery from Data* , 1: 1, 2007.

[37] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-

L.Barabási. The large-scale organization of metabolic networks. *Nature* , 407: 651-655, 2000.

[38] S. Dorogovtsev and J. Mendes. Effect of the accelerating growth of communications networks on their structure. *Phys. Rev. E* , 63: 025101(R), 2001.

[39] M. J. Gagen and J. S. Mattick. Accelerating, hyperaccelerating, and decelerating networks. *Phys. Rev. E* , 72:016123, 2005.

[40] C. Cooper and P. Prałat. Scale-free graphs of increasing degree. *Random Structures & Algorithms* , 38: 396-421, 2011.

[41] N. Deo and A. Cami. Preferential deletion in dynamic models of web-like networks. *Inf. Proc. Lett* ., 102: 156-162, 2007.

[42] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks with aging of sites. *Phys. Rev. E* , 62: 1842, 2000.

[43] A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proc. National Academy of Sciences USA* , 97: 11149, 2000.

[44] K. Klemm and V. M. Eguiluz. Highly clustered scale-free networks. *Phys. Rev. E* , 65: 036123, 2002.

[45] X. Zhu, R. Wang, and J.-Y. Zhu. The effect of aging on network structure. *Phys. Rev. E* , 68: 056121, 2003.

第6章 度相关性

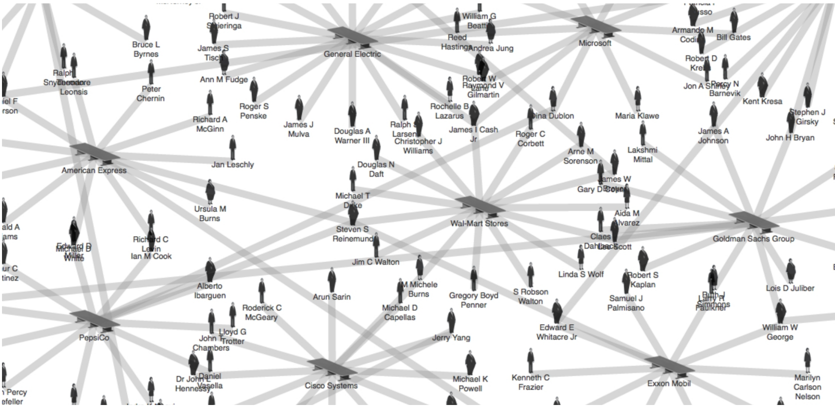


图6-0　TheyRule.net，作者乔希·翁

交互式网站TheyRule.net由旧金山设计师乔希·翁（Josh On）创建，使用网络来表达美国经济环环相扣的关系，展示了就职于多家美国顶级公司董事会的商业人士，揭示了这一小部分人的影响力。自2001年发表以来，这一项目被认为兼具艺术性和科学性。

本章索引

6.1 本章概览 / **287**

6.2 同配性和异配性 / **290**

6.3 计算度相关性 / **294**

6.4 结构截断 / **298**

6.5 真实网络的相关性 / **304**

6.6 生成相关网络 / **309**

6.7 度相关性的影响 / **316**

6.8 小结 / **319**

6.9 课后习题 / **322**

6.10 进阶阅读**6.A**
度相关性系数 / **324**

6.11 进阶阅读**6.B**
结构截断 / **328**

6.12 参考文献 / **330**

本章致谢

马顿·波什福伊 **MÁRTON PÓSFAI**

罗伯塔·西纳特拉 **ROBERTA SINATRA**

加布里埃莱·穆塞拉 **GABRIELE MUSELLA**

萨拉·莫里森 **SARAH MORRISON**

莫罗·马蒂诺　**MAURO MARTINO**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

菲利普·赫费尔　**PHILIPP HOEVEL**

妮科尔·萨迈　**NICOLE SAMAY**

6.1
本章概览

安吉丽娜·朱莉（Angelina Jolie）和布拉德·皮特（Brad Pitt）、本·阿弗莱克（Ben Affleck）和詹妮弗·加纳（Jennifer Garner）、哈里森·福特（Harrison Ford）和卡莉斯塔·弗洛克哈特（Calista Flockhart）、迈克尔·道格拉斯（Michael Douglas）和凯瑟琳·泽塔-琼斯（Catherine Zeta-Jones）、汤姆·克鲁斯（Tom Cruise）和凯蒂·霍尔姆斯（Katie Holmes）、理查·基尔（Richard Gere）和辛迪·克劳馥（Cindy Crawford）（图**6-1**），这是个有意思的名单。如果你经常看新闻标题，就会一眼认出名单里的人都是现在或曾经结婚的好莱坞明星。他们的婚礼（以及离婚）引发了不计其数的媒体报道，使八卦杂志卖出了数百万份。多亏了他们，人们才相信明星会与明星结婚，但我们很少停下来问一句：这合理吗？换言之，一个明星与另一个明星结婚的真实概率是多大？



图**6-1** 枢纽节点与枢纽节点约会

明星夫妻是社交网络中枢纽节点倾向于相互认识、约会和结婚的佐证。

假设一个明星可以从包含全世界1亿（10^8）适龄个体的池子里任选一个人来约会，则其伴侣也是一个明星（假设一共有1 000个明星）的概率只有10^{-5}。因此，如果约会是由随机邂逅所致，明星几乎不会与明星结婚。

虽然我们并不关心明星的约会习惯，但也得停下来看看，这一有趣的现象反映了社交网络结构的什么特性。明星、政治领袖、大型公司的首席执行官们往往认识非常多的人，且有更多人认识他们，这些人都是枢纽节点。因此明星间约会（图**6-1**）、名人担任多个公司董

事（图6-0　）等事件都是社交网络中一个有趣特性的表现形式。这个特性就是：枢纽节点倾向于与其他枢纽节点相连。

当然，并非所有网络都拥有这一性质，这似乎是显而易见的。以图6-2　所示的酵母蛋白质相互作用网络为例，显然该网络是一个无标度网络：大量的度为1或2的蛋白质和少数拥有很多链接的枢纽节点共存。然而这些枢纽节点倾向于彼此不相连，而是各自连接到大量小度节点，使网络形成中心辐射模式。这一现象在图6-2　中高亮的两个枢纽节点上表现得尤为明显：它们几乎只与度很小的蛋白质相互作用。

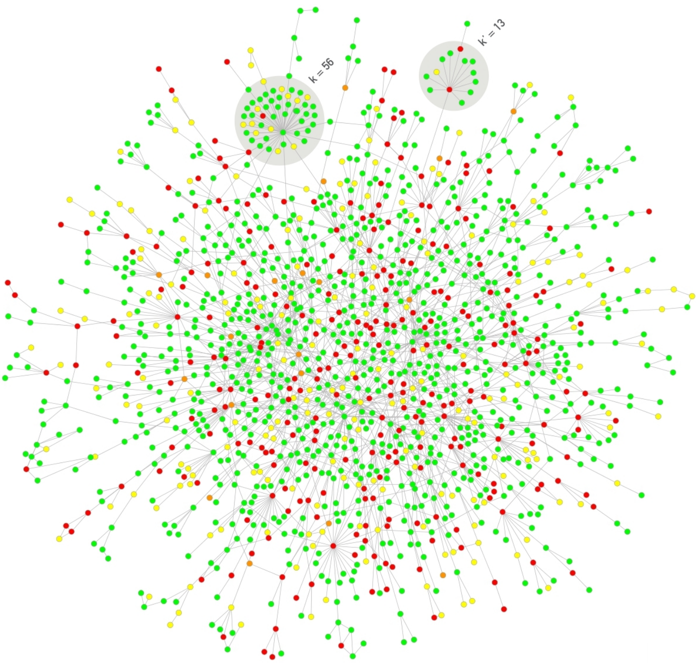


图6-2　枢纽节点彼此不相连

酵母的蛋白质相互作用网络。每个节点对应一种蛋白质，两种蛋白质相连表示有实验证据表明这两种蛋白质可以在细胞内结合。我们高亮显示了最大的两个枢纽节点，度分别为$k=56$和$k'=13$。这两个节点都与大量的小度节点相连，但彼此并未相连。

该网络有$N=1\ 870$种蛋白质和$L=2\ 277$条链接，是最早的蛋白质相互作用网络之一[1]，[2]。此处只显示了巨连通分支，表3-1 中的酵母蛋白质相互作用网络是后来绘制的，比本图有更多的节点和链接。节点颜色表示每种蛋白质的必要性：红色节点被移

除会导致生物体死亡，因此这些节点被称为致命蛋白质或必需蛋白质。与之相反，移除一个绿色节点后，生物体仍能存活。[3]

我们通过一个简单的计算来说明这种模式有多么不同寻常。假设每个节点随机选择其他节点相连，则度为k和k'的两个节点相连的概率是：

$$p_{k,k'}=\frac{kk'}{2L} \qquad (6.1)$$

公式**6.1** 告诉我们，枢纽节点因具有大量链接的属性，更可能彼此相连，而非连接到小度节点上。事实上，如果k和k'都很大，$p_{k,k'}$也很大。因此，度为$k=56$和$k'=13$的两个枢纽节点之间存在链接的概率是$p_{k,k'}=0.16$，比一个度为2的节点连接到一个度为1的节点的概率$p_{1,2}=0.000\ 4$高400倍。然而，图**6-2** 中的两个枢纽节点之间并没有链接，但我们却观测到有大量链接存在于度很小的节点之间。

图**6-2** 中高亮的两个枢纽节点几乎只连接到度为1的节点。当然，枢纽节点连接到度为1的节点并不意外：一个度为$k=56$的枢纽节点应该连接到$N_1 p_{1,56}\approx 12$个$k=1$的节点。但是，这个枢纽节点却拥有46个度为1的邻居，大约为期望值的4倍。

总而言之，虽然社交网络的枢纽节点倾向于彼此“约会”，蛋白质相互作用网络的情况则相反：枢纽节点倾向于彼此不相连，而是连接大量的小度节点。虽然仅通过两个例子便得出一般规律是危险的，但本章后面的部分将说明，正是这些模式展示出真实网络的普遍属性：一种名为度相关性的现象。接下来我们将讨论如何度量度相关性，并探讨其对网络拓扑结构的影响。

6.2 同配性和异配性

因拥有大量链接，枢纽节点本应彼此相连，在有些网络中是这样，但也存在一些例外。如图**6-3** 所示，三个网络有相同的度序列，

却有不同的拓扑：

- 中性网络（**Neutral Network**）

图**6-3b** 是一个随机网络。我们称这个网络是中性的，枢纽节点之间的链接数量正好与公式**6.1** 的随机预测相吻合。

- 同配网络（**Assortative Network**）

图**6-3a** 的网络度序列与图**6-3b** 的网络度序列完全一致。然而，图**6-3a** 的枢纽节点倾向于彼此相连，而非连接到小度节点，同时小度节点倾向于连接到其他小度节点。表现出这一趋势的网络被称为同配网络。这类模式的一个极端表现是完全同配网络，即度为k 的节点只连接到度为k 的节点（图**6-4** ）。

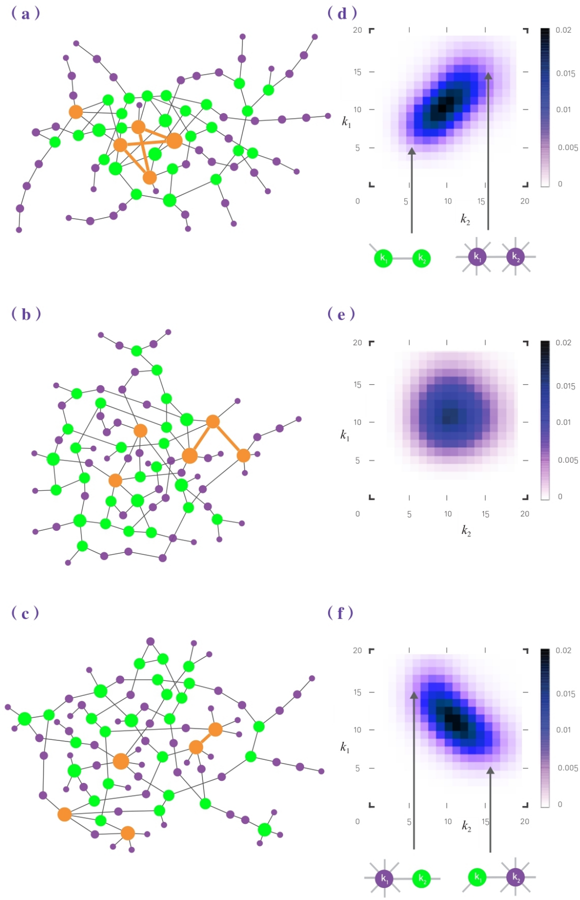


图6-3　度相关性矩阵

（a）（b）（c）　度分布完全一致的3个网络，但它们的度相关性不同。图中只展示了最大连通分支，橙色高亮显示度最大的5个节点及它们之间的链接。

（d）（e）（f） 度相关性矩阵e_{ij} ，分别对应同配网络（d）、中性网络（e）和异配网络（f）。各网络的度分布都为泊松分布，$N=1\ 000$且$\langle k\rangle=10$。颜色深浅表示随机选择一条边，该边连接度为k_1和k_2的两个节点的概率。

（a）（d）同配网络

同配网络的e_{ij}　在主对角线上较高。这表明度相近的节点倾向于彼此相连：小度节点连接小度节点，枢纽节点连接枢纽节点。事实上，图a中的网络有大量链接存在于枢纽节点之间，也有大量链接存在于小度节点之间。

（b）（e）中性网络

中性网络的节点连接是随机的。因此链接密度关于平均度对称，表明其连接模式没有相关性。

（c）（f）异配网络

异配网络的e_{ij} 在次对角线上较高，表明枢纽节点倾向于与小度节点相连，小度节点倾向于与枢纽节点相连。因此，这些网络有中心辐射特征，如图c所示。

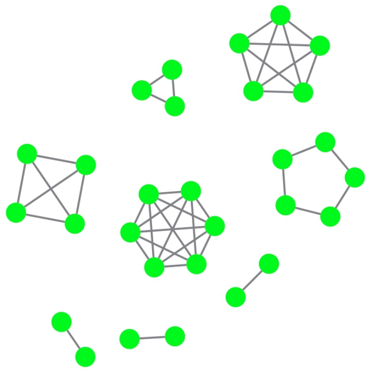


图6-4　完全同配

在一个完全同配网络中，每个节点只连接到度相同的其他节点。因此$e_{jk}=\delta_{jk}q_k$，其中δ_{jk}是克罗内克函数（Kronecker delta）。此时，矩阵e_{jk}的所有非对角元素为零。图中展示了一个完全同配网络，完全由k-团构成。

- 异配网络（**Disassortative Network**）

图**6-3c** 中的枢纽节点倾向于彼此不相连，而是连接到小度节点。因此，网络表现出中心辐射特征，这种模式被称为异配。

通常来说，当网络中大度节点和小度节点之间的链接数量系统性地异于随机期望时，网络呈现出度相关性。此处随机期望为公式**6.1**推导出的度k节点和度k'节点之间的链接数量。

度相关性矩阵可用于表达网络的度相关性信息，其中每个元素e_{ij}表示随机选择一条链接，其两端点的度分别为i和j的概率。e_{ij}既然是概率，则必然是归一的，即

$$\sum_{i,j} e_{ij}=1 \qquad (6.2)$$

根据公式4.27 推导的概率q_k ，即随机选择的一条链接有一个端点度为k 的概率，可知

$$q_k = \frac{kp_k}{\langle k \rangle} \tag{6.3}$$

通过下式，q_k 与e_{ij} 可以联系起来：

$$\sum_j e_{ij} = q_i \tag{6.4}$$

对于随机连接的中性网络，我们期望

$$e_{ij} = q_i q_j \tag{6.5}$$

那么当e_{ij} 不同于随机期望公式**6.5** 时，网络便具有度相关性。注意，公式**6.2** ~**6.5** 适用于具有任何度分布的网络，因此随机网络和无标度网络都可以这样计算。

e_{ij} 包含了度相关性的所有信息，我们可以将其可视化，图**6-3d**、**e**、**f** 展示了同配网络、中性网络和异配网络的e_{ij} 。在中性网络中，小度节点和大度节点随机相连，因此e_{ij} 不呈现任何趋势（图**6-3e** ）。在同配网络中，主对角线方向上存在高相关性，表明节点主要连接到度接近的节点，即小度节点倾向于连接到其他小度节点，枢纽节点倾向于连接到其他枢纽节点（图**6-3d** ）。与之相反，在异配网络中，e_{ij} 在次对角线上有高相关性，即大度节点倾向于连接到小度节点（图**6-3f** ）。

总而言之，度相关性矩阵包含了丰富的度相关性信息，但是通过观察e_{ij} 来研究度相关性有诸多缺点：

- 很难从矩阵的可视化表达中提取信息；
- 无法定量估计度相关性，从而很难在度相关性不同的网络中进行比较；
- 度相关性矩阵包含大约k_{max}^2 /2个独立变量，其中蕴含了大量信息，使我们难以通过解析计算和模拟来建模。

因此，我们需要一种更简洁的方法来定义度相关性，这正是接下来几节的目标。

6.3 计算度相关性

虽然e_{ij} 包含了一个特定网络中度相关性的全部信息，我们却很难对其进行解读。这一节将介绍度相关性函数，从而提供一种简洁的方法来量化度相关性。

度相关性刻画了相连节点的度之间的关系，一种量化方式是计算每个节点i 的邻居的平均度（图**6-5**）：

$$k_{nn}(k_i) = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}k_j \qquad (6.6)$$

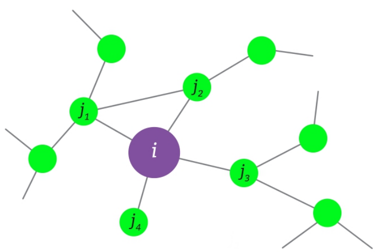


图**6-5** 最近邻的度：$\boldsymbol{k}_{nn}$

为了理解度相关性函数$k_{nn}(k_i)$，我们需要计算一个节点的邻居的平均度。图中展示了对节点i 计算$k_{nn}(k_i)$的过程。节点i 的度为$k_i=4$，计算其邻居j_1、j_2、j_3和j_4的平均度，可得$k_{nn}(4)=(4+3+3+1)/4=2.75$。

对所有度为k 的节点计算度相关性函数（公式**6.6**）[4]，[5]

$$k_{nn}(k) = \sum_{k'} k'P(k' \mid k) \qquad (6.7)$$

其中，$P(k'|k)$指对于一个度为k 的节点，存在一个度为k'的节点与之相连的条件概率，因此$k_{nn}(k)$是所有度为k 的节点的邻居平均

度。为了深入分析度相关性，我们将考察$k_{\mathrm{nn}}(k)$与k 的关系。

- 中性网络

对于中性网络，公式**6.3** ~**6.5** 断言

$$P(k' \mid k) = \frac{e_{kk'}}{\sum_{k'} e_{kk'}} = \frac{e_{kk'}}{q_k} = \frac{q_{k'} q_k}{q_k} = q_{k'} \qquad (6.8)$$

这允许我们将$k_{\mathrm{nn}}(k)$改写为

$$k_{\mathrm{nn}}(k) = \sum_{k'} k' q_{k'} = \sum_{k'} k' \frac{k' p(k')}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} \qquad (6.9)$$

因此，在中性网络中，一个节点的邻居的平均度独立于该节点的度k ，只取决于网络的全局特征$\langle k \rangle$ 和$\langle k^2 \rangle$ 。所以$k_{\mathrm{nn}}(k)$关于k 的曲线应该是水平线$\langle k^2 \rangle / \langle k \rangle$ ，这正是我们在电网网络中所观察到的（图**6-6b** ）。公式**6.9** 也刻画了真实网络中一个耐人寻味的性质：我们的朋友比我们自己更受欢迎，这一现象被称为朋友悖论（边栏**6.1** ）。

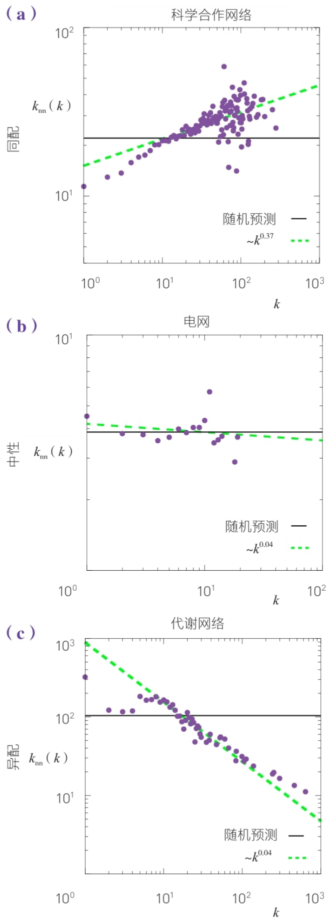


图6-6　度相关性函数

三个真实网络的度相关性函数$k_{nn}(k)$。各个子图均在双对数坐标下展示$k_{nn}(k)$，以检验其是否服从幂律（公式6.10）。

（a）科学合作网络

$k_{nn}(k)$随k递增，表明网络的同配性。

（b）电网

$k_{nn}(k)$是水平的，表明没有度相关性，符合中性网络的判断（公式6.9）。

（c）代谢网络

$k_{nn}(k)$随k递减，表明网络的异配性。

每个子图的水平线对应公式6.9的判断，绿色虚线是对公式6.10的拟合。

边栏6.1

朋友悖谬

朋友悖谬是一个出乎意料的表述：平均而言，我的朋友比我更受欢迎[6]，[7]。这一论断来自公式6.9：一个节点的邻居的平均度不是简单的$\langle k\rangle$，而是也与$\langle k^2\rangle$有关。

考虑一个随机网络，有$\langle k^2\rangle = \langle k\rangle(1+\langle k\rangle)$。根据公式6.9可知$k_{nn}(k)=1+\langle k\rangle$。因此，一个节点的邻居的平均度总是高于随机选择节点的平均度$\langle k\rangle$。

在无标度网络中，$\langle k\rangle$与我们朋友的度可能相差很大，因为$\langle k^2\rangle/\langle k\rangle$远远大于$\langle k\rangle$（图3-8）。以演员网络为例，有$\langle k^2\rangle/\langle k\rangle=565$（表3-1）。一个节点的朋友的平均度是该节点自己的度的数百倍。

朋友悖谬的来源很简单：我们更可能与枢纽节点而非小度节点做朋友，仅仅是因为枢纽节点比小度节点有更多的朋友。

- 同配网络

在同配网络中，枢纽节点倾向于连接到其他枢纽节点，于是节点的度k越高，其最近邻居的平均度也越高。因此，对于同配网络，$k_{nn}(k)$随k递增，正如我们在科学合作网络中所观察到的（图**6-6a**）。

- 异配网络

在异配网络中，枢纽节点倾向于连接到低度节点。因此$k_{nn}(k)$随k递减，正如我们在代谢网络中所观察到的（图**6-6c**）。

图**6-6**中观察到的现象启发我们利用下式来近似度相关性函数[4]

$$k_{nn}(k) = ak^{\mu} \tag{6.10}$$

若公式6.10成立，则度相关性本质上由相关性指数μ的符号决

定：

- 同配网络：$\mu > 0$

通过拟合科学合作网络的$k_{nn}(k)$可得$\mu = 0.37 \pm 0.11$（图**6-6a**）。

- 中性网络：$\mu = 0$

根据公式**6.9**，$k_{nn}(k)$独立于k。事实上，拟合电网网络可得$\mu = 0.04 \pm 0.05$，这个值可以近似为零（图**6-6b**）。

- 异配网络：$\mu < 0$

通过拟合代谢网络可得$\mu = -0.76 \pm 0.04$（图**6-6c**）。

总而言之，度相关性函数可以帮助我们刻画真实网络中相关性存在与否。另外，$k_{nn}(k)$函数在解析阶段也扮演了重要角色，使得我们可以预测度相关性对各种网络特征的影响（**6.6**节）。然而，如果能用一个数字来刻画网络中相关性的大小则更为方便：公式**6.10**中定义的相关性指数μ，或者边栏**6.2**介绍的度相关性系数可以达到这个目的。

边栏6.2

度相关性系数

如果希望使用单个数字来描述度相关性，我们可以使用μ或者度相关性系数。度相关性系数由马克·纽曼提出[8]，[9]，定义为：

$$r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\sigma^2} \quad (6.11)$$

且

$$\sigma^2 = \sum_k k^2 q_k - \left[\sum_k k q_k\right]^2 \quad (6.12)$$

则r 是一条链接的两个端点的度的皮尔逊相关系数，取值范围为-1≤r ≤1：r >0时，网络是同配的；r =0时，网络是中性的；r <0时，网络是异配的。例如，科学合作网络的r =0.13，蛋白质相互作用网络的r =-0.04，电网的r =0，均符合此前所述的同配/异配性。

度相关性系数背后的假设是k_{nn} (k)以斜率r 线性相关于k 。与之相反，相关性指数μ 假设k_{nn} (k)服从幂律**6.10** 。这两个假设自然不可能同时成立。**6.7**节 的分析提供了一些指导，支持公式**6.10**成立，另外，如进阶阅读**6.A** 中所示，通常r 与μ 相关。

6.4 结构截断

本书中，我们假设网络是简单网络，即两个节点之间最多存在一条链接（图**1-17** ）。例如在电子邮件网络中，有邮件往来的两个个体之间只有一条链接，即使他们可能有多次邮件往来。与之类似，在演员网络中，共同出演过同一部电影的两个演员之间只有一条链接，无论他们共同出演过多少部电影。表**3-1** 列出的所有数据集都是简单网络。

在简单网络中，无标度属性和度相关性存在一个令人费解的矛盾[10]，[11] 。例如图**6-7a** 中的无标度网络，度最大的两个枢纽节点的度为k =55和k ′=46。若网络中存在度相关性$e_{k,k'}$，则这两个节点之间的期望链接数为

$$E_{kk'} = e_{kk'}\langle k\rangle N \quad (6.13)$$

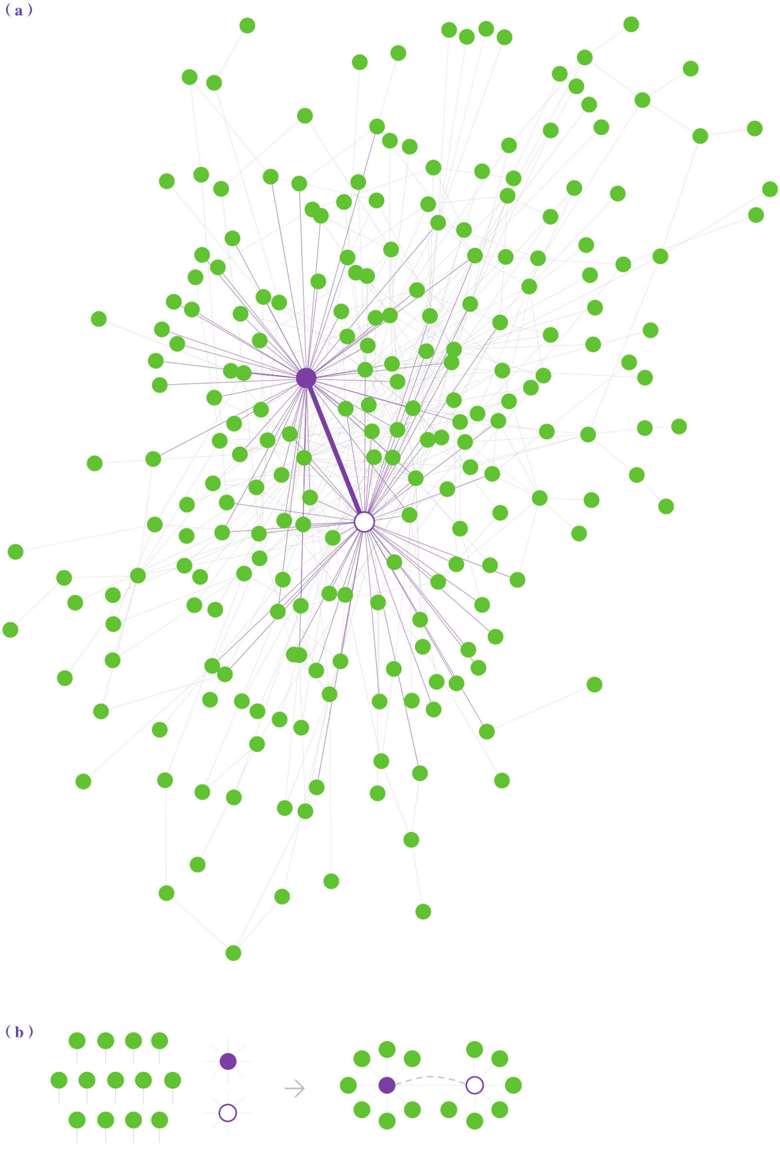


图6-7　结构性异配

（a）采用配置模型（图3-15　　）生成的无标度网络，N＝300，L　　＝450且γ＝2.2，我们通过禁止自环和多重链接，使其变成简单网络。图中高亮显示了网络中最大的两个节点。如公式**6.14**　所示，为了使网络保持中性，这两个节点之间需要有大约三条链接。然而，我们禁止多重链接的要求（简单网络约束）使得这个网络是异配的，这种现象被称为结构性异配。

（b）我们从一个固定的度序列开始展示结构性相关性的来源。如左图中仅带有一半链接的节点所示，接下来我们将这些半链接随机相连（配置模型）。则度为8和7的两个节点之间的链接数的期望值是$8\times7/28 \approx 2$。然而，如果我们不允许多重链接存在，则这两个节点之间只可能有一条链接，从而使这个网络结构性异配。

中性网络的$e_{k,k'}$由公式**6.5** 给出，结合公式**6.3** 可得

$$E_{kk'}=\frac{kp_k k'p_{k'}}{\langle k\rangle}N=\frac{\frac{55}{300}\cdot\frac{46}{300}}{3}300=2.8 \qquad (6.14)$$

因此，对于这两个大度枢纽节点，若网络是中性的，则它们之间应该有2到3条链接。然而，在简单网络中，两个节点之间只能有一条链接，这就导致了度相关性与无标度属性的矛盾。本节的目标是理解这一矛盾的起源和结果。

当k 和k'较小时，公式**6.14** 表明$E_{k,k'}$也较小，也就是说，两个节点之间的链接应该少于一条。只有当节点的度高于某个阈值k_s 时，公式**6.14** 才表明会有多条链接。如进阶阅读**6.B** 所示，k_s 被称为结构截断，满足下式：

$$k_s(N)\sim(\langle k\rangle N)^{1/2} \qquad (6.15)$$

换言之，若节点的度超过公式**6.15** ，则有$E_{k,k'}>1$。接下来我们将说明这一矛盾导致了度相关性。

要理解结构截断，首先要通过比较结构截断k_s 和自然截断k_{max} ，来看一个网络中是否有节点的度超过公式**6.15** ，其中k_{max} 是网络中最大度的期望。根据公式**3.18** ，无标度网络中$k_{max}\sim N$ ，根据k_s 和k_{max} 的大小关系可以分为两类网络：

- 无结构截断

对于随机网络和$\gamma\geq3$的无标度网络，k_{max} 的指数小于1/2，因此k_{max} 总是小于k_s 。换言之，结构截断所需的度高于最大枢纽节点的度，因此没有节点存在$E_{k,k'}>1$。这类网络中，度相关性和简单网络条件没有矛盾。

- 结构性异配

对于$\gamma<3$的无标度网络，我们有$1/(\gamma-1)>1/2$，即k_s可以小于k_{max}。因此，度介于k_s和k_{max}之间的节点将导致$E_{k,k'}>1$。换言之，网络中枢纽节点之间的链接数少于公式**6.14**的期望，因此这类网络将表现出异配性，我们称这一现象为结构性异配。如图**6-8a**、**b**中由配置模型生成的简单无标度网络所示，尽管我们并未在其生成过程中引入度相关性，网络仍表现出异配性。

有两种方法可以生成没有结构性异配现象的网络：

（**1**） 放宽简单网络条件，允许节点之间有多条链接。从而矛盾消失，网络将表现出中性（图**6-8c**、**d**）。

（**2**） 如果我们坚持要有简单无标度的中性或同配网络，就必须移除度大于k_s的所有枢纽节点。如图**6-8e**、**f**所示，若网络中没有$k\geq 100$的节点，则网络是中性的。

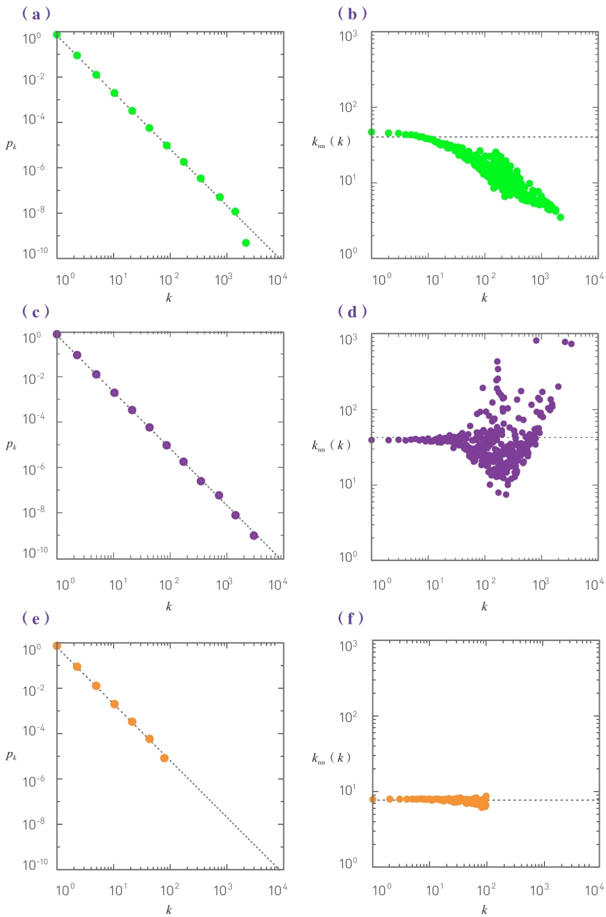


图6-8　自然截断和结构截断

图中展示了无标度属性和度相关性之间的联系。这是由配置模型（图**3-15**　）生成的$N=10\ 000$且$\gamma=2.5$的无标度网络的度分布（左侧子图）和度相关性函数$k_{nn}(k)$（右侧子图）。

（a）（b）　如果生成的无标度网络具有图a中的幂律度分布，且没有自环和多重链接，则这个网络表现出结构性异配，如图b中的$k_{nn}(k)$所示。在这种情况下，度较大的节点之间没有足够多的链接来维持网络中性，因此对于较高的k来说，$k_{nn}(k)$一定会下降。

（c）（d）　我们可以放宽简单网络条件来消除结构性异配，即允许两个节点之间有多重链接。如图c和图d所示，这种情况下我们会得到中性的无标度网络。

（e）（f）　如果我们移除所有$k \geq k_s \simeq 100$的节点，从而设置高位截断（upper cutoff），如公式**6.15** 所示，网络将变成图f所示的中性网络。

但是如何知道某个网络中观察到的相关性是由结构性异配所致，还是由其他未知的导致度相关性的原因所致？保度随机化（图**3-17**）可以区分这两种可能性：

（**1**）简单链接的保度随机化（**R-S**）

对原始网络做保度随机化。每一步，我们不允许两个节点之间有多于一条链接。从算法角度来看，这意味着若一次连接可能生成多重链接，则要取消这次连接。如果真实的$k_{nn}(k)$和随机化之后的$k_{nn}^{R-S}(k)$不可区分，那么真实系统中观察到的相关性是结构性的，可完全由度分布解释。如果随机化之后的$k_{nn}^{R-S}(k)$没有表现出$k_{nn}(k)$表现出的度相关性，则存在某些未知的导致度相关性的过程。

（**2**）多重链接的保度随机化（**R-M**）

作为一种自洽检查，允许节点间存在多重链接的保度随机化在有些时候是有用的。从算法角度来看，这意味着我们总是允许随机重连，无论是否产生了多重链接。这一过程会消除所有度相关性。

我们对三个真实网络进行上述随机化。如图**6-9a** 所示，经过这两种随机化后，科学合作网络的同配性都消失了。这表明合作网络的同配相关性并非来自其无标度属性。相反，代谢网络中观察到的异配性在R-S随机化后保持不变（图**6-9c**）。因此，代谢网络的异配性是结构性的，由它的度分布所致。

总而言之，无标度属性可以导致简单网络中的异配性。事实上，在中性网络或同配网络中，枢纽节点之间应该有多重链接。如果不允许多重链接存在（简单图），网络将表现出异配倾向。随机网络和$\gamma \geq 3$的无标度网络不存在这一矛盾。若我们允许节点之间有多重链接，这一矛盾也会消失。

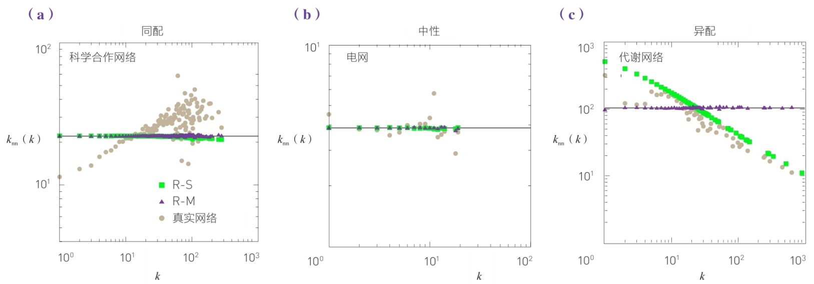


图6-9　随机化和度相关性

为了找到度相关性的来源，我们必须比较$k_{nn}(k)$（图中灰色）与保度随机化后的$k_{nn}^{R-S}(k)$和$k_{nn}^{R-M}(k)$。在这一过程中，这两种保度随机化方法都可以提供丰富的信息：

简单链接的保度随机化（**R-S**）：随机化的每一步都要检查确认，任意两个节点之间都没有多重链接。

多重链接的保度随机化（**R-M**）：随机化过程中允许多重链接存在。

我们对图**6-6**中的网络进行这两种随机化。R-M过程总是生成中性网络，因此$k_{nn}^{R-M}(k)$总是水平的。比较$k_{nn}(k)$和$k_{nn}^{R-S}(k)$可以帮助我们确定所观察到的相关性是不是结构性的：

（**a**）科学合作网络

$k_{nn}(k)$递增，不同于水平的$k_{nn}^{R-S}(k)$，表明网络的同配性不是结构性的，因此，这一同配性来自该网络演化中的某些过程。这并不意外：结构性的影响只会产生异配性，不会产生同配性。

（**b**）电网

$k_{nn}(k)$、$k_{nn}^{R-S}(k)$和$k_{nn}^{R-M}(k)$均保持水平，表明网络中不存在度相关性（中性网络）。

（**c**）代谢网络

$k_{nn}(k)$和$k_{nn}^{R-S}(k)$都是递减的，我们认为该网络的异配性来自其自身的无标度属性。因此，观察到的度相关性是结构性的。

6.5
真实网络的相关性

为了理解度相关性的普遍性，我们查看了真实网络中的度相关性。图**6-10**展示了在10个真实网络中，$k_{nn}(k)$函数具有以下几种不

同的模式：

- 电网

电网的$k_{\mathrm{nn}}(k)$是水平的，且与其随机化后的值非常相近，表明它没有度相关性（图**6-10a**），因此电网是中性的。

- 互联网

小度节点（$k \leq 30$）的$k_{\mathrm{nn}}(k)$表现出清晰的同配趋势，大度节点则没有这一趋势（图**6-10b**）。随机化后，度相关性消失，因此互联网是同配的，但结构截断让大度节点没有这一趋势。

- 社交网络

刻画社会交互行为的三个网络——手机通话网络、科学合作网络和演员网络都有递增的$k_{\mathrm{nn}}(k)$，它们是同配网络（图**6-10c**～**e**）。因此，这些网络的枢纽节点倾向于连接到其他枢纽节点，小度节点倾向于连接到小度节点。$k_{\mathrm{nn}}(k)$与$k_{\mathrm{nn}}^{\mathrm{R-S}}(k)$不同，表明社交网络的同配性并非由它们无标度的度分布所致。

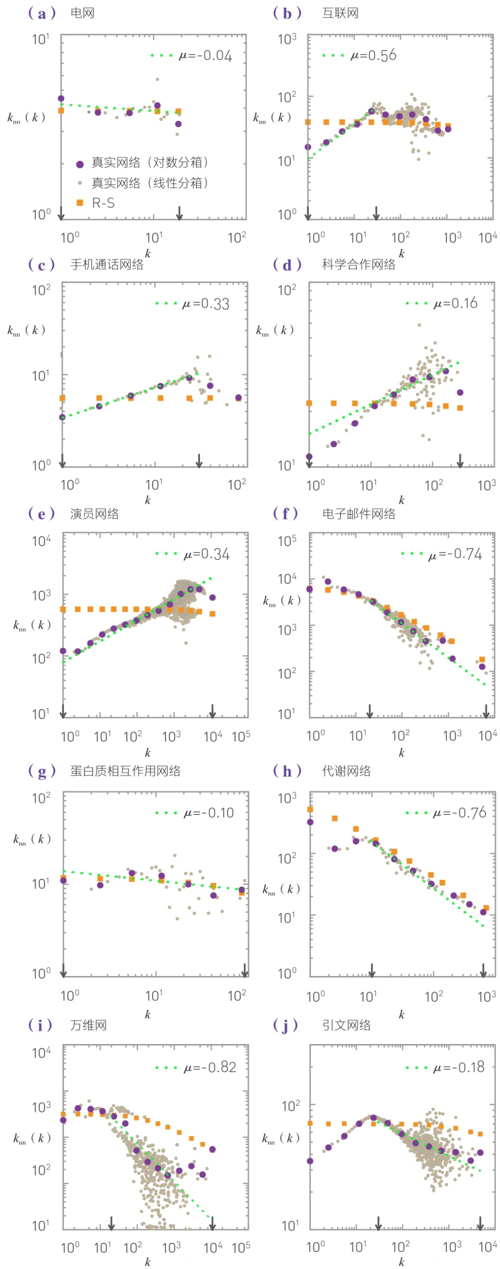


图6-10　随机化和度相关性

10个真实网络（表3-1　）的度相关性函数$k_{nn}(k)$。灰色记号为使用线性函数表示

的$k_{nn}(k)$函数，紫色圆圈代表使用对数拟合的同一数据（**3.11**节）。绿色虚线为公式**6.10**的最优拟合，拟合区域如底部箭头所示。橙色方块表示100次独立的、保持网络基本属性的保度随机化所得的$k_{nn}^{R-S}(k)$。注意，计算$k_{nn}(k)$时，我们将有向网络视为无向网络。要完全刻画有向网络中产生的相关性，应使用有向相关性函数（边栏**6.3**）。

边栏**6.3**

有向网络的相关性

度相关性函数**6.7**是定义在无向网络上的。要度量有向网络上的相关性，我们需要考虑到每个节点i的特征是由入度k_i^{in}和出度k_i^{out}共同表征的[14]。因此我们定义4个度相关性函数$k_{nn}^{\alpha,\beta}(k)$，其中α和β分别表示入度（in）和出度（out）（图**6-11a**～**d**）。图**6-11e**展示了引文网络的$k_{nn}^{\alpha,\beta}(k)$，表明入度 - 出度相关性不存在，而其他三种相关性（入度 - 入度，出度 - 入度，出度 - 出度）对较小的k存在同配性。

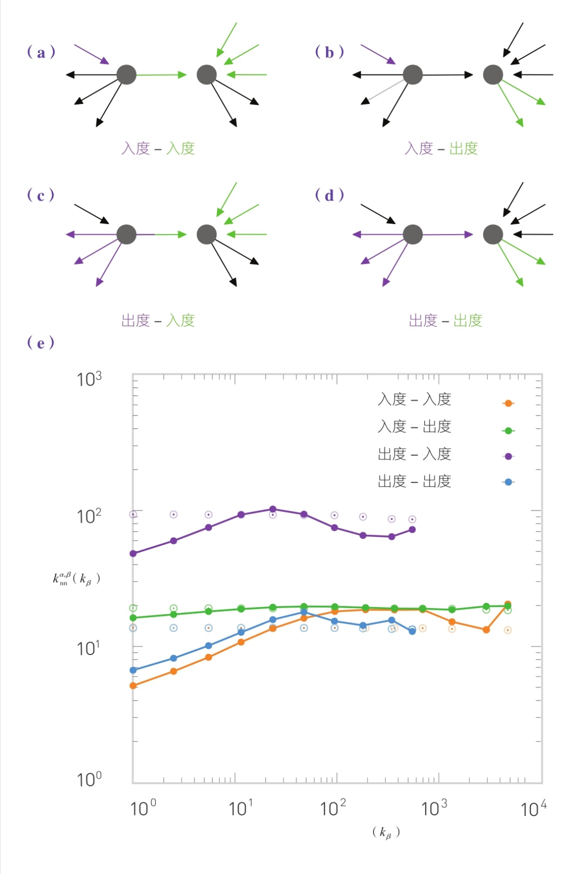


图6-11　有向网络的相关性

（a）～（d）刻画有向网络的4种可能的相关性。我们用紫色和绿色表示(α，β)，以定义各个相关性函数[14]。例如，图a给出了$k_{\mathrm{nn}}^{\mathrm{in,in}}(k)$相关性，即两个相邻节点的入度的相关性。

（e）有向的引文网络的$k_{\mathrm{nn}}^{\alpha,\beta}(k)$相关性函数。例如，$k_{\mathrm{nn}}^{\mathrm{in,in}}(k)$是入度为$k_{\mathrm{in}}$的节点连入邻居的平均入度。4个函数中有3个在度小于$k \lesssim 100$时表现出清晰的同配倾向。空心记号表示对于各个度相关性函数进行保度随机化后的$k_{\mathrm{nn}}^{\alpha,\beta}(k)$（R-S随机化）。

- ## 电子邮件网络

虽然电子邮件网络通常被视为一种社交网络，但它的$k_{\mathrm{nn}}(k)$随k递减，表现出清晰的异配性（图6-10f）。随机化后的$k_{\mathrm{nn}}^{\mathrm{R-S}}(k)$也是

递减的，表明我们观察到的是结构性异配，其来源是网络的无标度属性。

- 生物网络

蛋白质相互作用网络和代谢网络的μ 都是负值，表明这些网络是异配网络。然而，$k_{nn}^{R-S}(k)$ 与$k_{nn}(k)$非常近似，表明我们观察到的是结构性异配，是由网络的无标度属性（图**6-10g**、**h**）导致的。

- 万维网

递减的$k_{nn}(k)$表明万维网具有异配相关性（图**6-10i**），随机化后的$k_{nn}^{R-S}(k)$ 也是递减的，但递减的速度比$k_{nn}(k)$慢。因此，万维网的度分布不能完全解释其异配性。

- 引文网络

这个网络有一个令人疑惑的现象：$k \le 20$时，度相关性函数$k_{nn}(k)$表现出清晰的同配趋势；然而当$k > 20$时，我们观察到异配性（图**6-10j**）。这一混合形态可以在极端同配的网络中找到（图**6-13b**）。这表明引文网络有很强的同配性，但其无标度属性导致了结构性异配，并改变了$k \gg k_s$ 时$k_{nn}(k)$的斜率。

总而言之，图**6-10** 表明，我们必须通过比较$k_{nn}(k)$和保度随机化后的$k_{nn}^{R-S}(k)$ 才能理解度相关性，这也让我们归纳出几个有趣的结论：

（**1**） 10个真实网络中，电网是唯一一个真正的中性网络。因此，大多数真实网络都存在度相关性。

（**2**） 所有表现出异配倾向的网络（电子邮件网络、蛋白质相互作用网络、代谢网络），其异配倾向都来源于其无标度属性。因此，它们都是结构性异配。只有万维网的异配相关性能够部分地由它的度分布解释。

（**3**） 同配网络的度相关性并不能由它的度分布解释。大多数社

交网络（手机通话网络、科学合作网络、演员网络）、互联网和引文网络都是如此。

科研人员提出了多种机制来解释观察到的这些同配性出现的原因。例如，个体倾向于构建社区（第**8**章　的主题），可能导致同配相关性 [12] 。与之类似，社会上有无数种机制——例如职业委员会等，把枢纽节点放在一起，从而强化了社交网络和职场网络的同配性。最后，同配性作为一种被充分研究的社会现象 [13] ，指出了个体倾向于连接到具有类似背景和特征的其他个体，因此度相近的个体更可能彼此认识。这种同配性或许也可以用来解释明星婚姻（图**6-1**　）。

6.6
生成相关网络

要探讨度相关性对各种网络特征的影响，首先我们需要理解此前讨论的各种网络模型特有的相关性，同样，设计出可以调节度相关性的网络生成算法也非常重要，但正如本节内容所示，由于无标度属性和度相关性矛盾的存在，设计出这种算法并非易事。

静态模型的度相关性

埃尔德什 - 雷尼模型

根据随机网络的定义，它属于中性网络。随机网络没有枢纽节点，也不具有结构性相关性。因此对于埃尔德什 - 雷尼网络，对于公式**6.9** 所定义的$k_{\mathrm{nn}}(k)$，对任意$\langle k\rangle$和N 均有$\mu=0$。

配置模型

配置模型（图**3-15**　）也是中性的，且与所选择的度分布p_k　无关。这是因为该模型允许自环和多重链接存在，因此，枢纽节点引发的任何矛盾都可以由它们之间的多重链接解决。然而，如果我们强制要求网络是简单网络，那么生成的网络将表现出结构性异配（图**6-8**

）。

隐参数模型

该模型的e_{jk} 正比于随机选择的隐变量η_j 和η_k 的乘积（图**3-18**）,因此该网络实际上是不相关的。然而，如果不允许多重链接存在，我们仍会在无标度网络上观测到结构性异配。在这种情况下，解析计算表明[18]：

$$k_{nn}(k) \sim k^{-1} \qquad (6.16)$$

即度相关性函数是公式**6.10** 的形式，且$\mu = -1$。

总的来看，此前讨论过的静态模型要么是中性网络，要么满足公式**6.16** 的结构性异配。

演化网络的度相关性

为了理解生长网络上度相关性的出现或消失，我们从初始吸引模型开始分析（**5.5**节 ）。巴拉巴西 - 阿尔伯特模型正是该类模型的一个特例。

初始吸引模型

考虑一个生长网络，其中存在公式**5.23** 的偏好连接，即$\Pi(k) \sim A + k$ ，其中A 是初始吸引因子。当A 取不同的值时，我们观察到三个区间，分别有不同的$k_{nn}(k)$[15]：

（**1**）异配区：$\gamma < 3$

若$-m < A < 0$，我们有

$$k_{nn}(k) \sim m\frac{(m+A)^{1-\frac{A}{m}}}{2m+A}\zeta\left(\frac{2m}{2m+A}\right)N^{\frac{A}{2m+A}}k^{\frac{A}{m}} \qquad (6.17)$$

因此该网络是异配的，$k_{nn}(k)$依幂律递减[15]，[16]：

$$k_{nn}(k) \sim k^{-\frac{|A|}{m}} \quad (6.18)$$

（2）中性区：$\gamma = 3$

若$A = 0$，初始吸引模型退化为巴拉巴西 - 阿尔伯特模型，则有

$$k_{nn}(k) \sim \frac{m}{2}\ln N \quad (6.19)$$

因此，$k_{nn}(k)$独立于k，网络是中性的。

（3）弱同配：$\gamma > 3$

若$A > 0$，计算得

$$k_{nn}(k) \approx (m+A)\ln\left(\frac{k}{m+A}\right) \quad (6.20)$$

由于$k_{nn}(k)$随k对数递增，该网络表现出弱同配趋势，但并不满足公式6.10。

总而言之，公式6.17 ~6.20 表明初始吸引模型生成了比较复杂的度相关性，从异配性到弱同配性都有可能。公式6.19 说明中性也有可能，比如巴拉巴西 - 阿尔伯特模型。最后，公式6.17 指出$k_{nn}(k)$与k存在对数依赖，为公式6.10 中$k_{nn}(k)$的近似方式提供了理论支持。

- 比安科尼 - 巴拉巴西模型

当适应性分布是均匀分布的时候，比安科尼 - 巴拉巴西模型生成的网络是异配的[5]（图6-12 ）。该网络随机化后依然是异配的，表明该模型的异配性是结构性的。但是，原网络的$k_{nn}(k)$和随机化后的$k_{nn}^{R-S}(k)$并不重合，说明该模型的异配性不能完全由其无标度属性来解释。

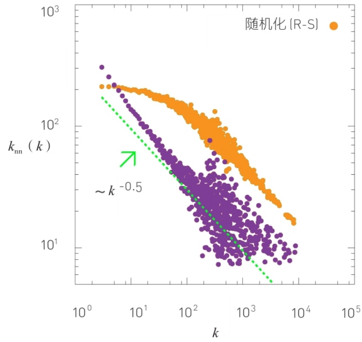


图6-12　比安科尼 - 巴拉巴西模型的相关性

比安科尼 - 巴拉巴西模型的度相关性函数，其中$N=10\ 000$，$m=3$，且有均匀的适应性分布（**5.2**节　）。根据公式**6.10**　拟合的绿色虚线可以看到，该网络是异配的，这与$\mu\backsimeq$ -0.5相合。橙色记号表示$k_{nn}^{R-S}(k)$，因为$k_{nn}^{R-S}(k)$也是递减的，所以观察到的异配性大体上是结构性异配。然而，$k_{nn}^{R-S}(k)$与$k_{nn}(k)$的差异表明，结构性因素不能完全解释所观察到的度相关性。

度相关性的调节

一些算法可以按照预设的度相关性生成网络 [8]，[17]，[18]。接下来我们讨论哈维 - 布鲁内特（Xalvi-Brunet）和索科洛夫（Sokolov）提出的算法的简化版，旨在根据预设的度序列生成最大相关性的网络 [19]，[20]，[21]。算法分以下几步（图**6-13a**　）：

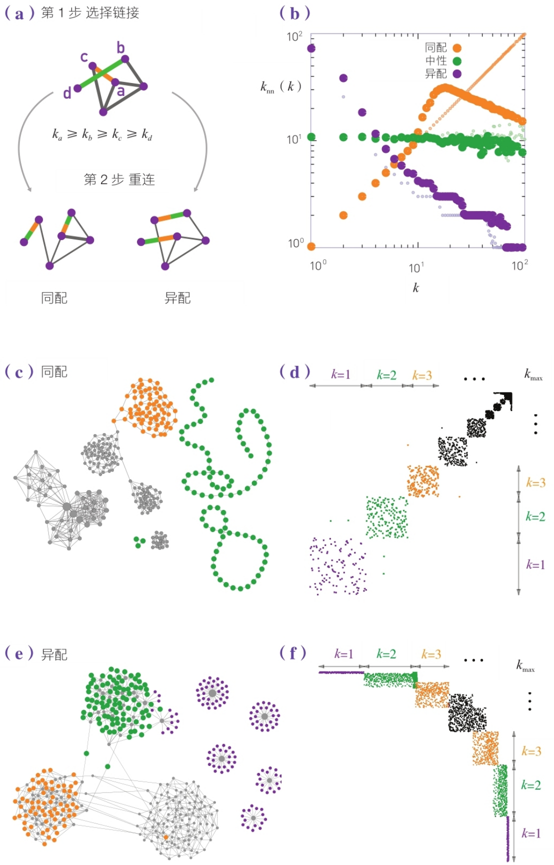


图6-13　布鲁内特和索科洛夫算法

该算法可以生成最大度相关性的网络。

（a）描述了算法的基本步骤。

（b）对于一个无标度网络，$N = 1\ 000$，$L = 2\ 500$，$\gamma = 3.0$，利用该算法生成网络的$k_{nn}(k)$。

（c）（d）利用算法生成的典型网络配置和相对应的最大同配网络的A_{ij}矩阵。其中A_{ij}的行和列按照节点度递增排序。

（e）（f）最大异配网络的网络配置和A_{ij}矩阵。

子图（d）和（f）的A_{ij}矩阵表达了最大相关性的网络的内在规律：网络由很多节点块组成，每个块的节点的度相近（图d）或相异（图f），节点与块内的其他节点相连。

- **第1步：选择链接**

随机选择两条链接，记4个端点为a、b、c、d，且这4个端点的度满足

$$k_a \geq k_b \geq k_c \geq k_d$$

- **第2步：重连**

删除选择的链接并重连，建立新的节点对。根据需要的度相关性，重连有两种方式：

- **第2A步**：同配

  将度较高的两个节点（a 和b ）配对，将度较低的两个节点（c 和d ）配对，从而让相连的两个节点度相近，增加了网络的同配性。

- **第2B步**：异配

  将度较高的两个节点和度较低的两个节点分别配对（a 和d，b 和c ），从而让节点与度不同的节点相连，增加了网络的异配性。

重复以上步骤，我们可以逐渐增加网络的同配性（第2A步）或异配性（第2B步）。如果希望生成简单网络（没有多重链接），在第2步之后，我们要检查这一重连是否会导致多重链接出现。如果出现了多重链接，则删除这一重连，回到第1步。

该算法生成的网络的相关性将收敛到给定度序列所能达到的最大值（同配）或最小值（异配）（图**6-13b**），该模型可以很容易地生成异配相关性（图**6-13e**、**f**）。在同配性达到极限的情况下，简单网络表现出混合的$k_{\mathrm{nn}}(k)$：k 较小时同配，k 较大时异配（图**6-13b**），这是结构截断的结果：对于无标度网络，系统无法对高k 保持同配性，这一现象类似于引文网络的$k_{\mathrm{nn}}(k)$函数（图**6-10j**）。

图**6-13** 介绍的这一版本的布鲁内特和索科洛夫算法可以生成最

大同配或最大异配网络。我们也可以使用图**6-14** 讨论的算法来调节生成网络度相关性的大小。

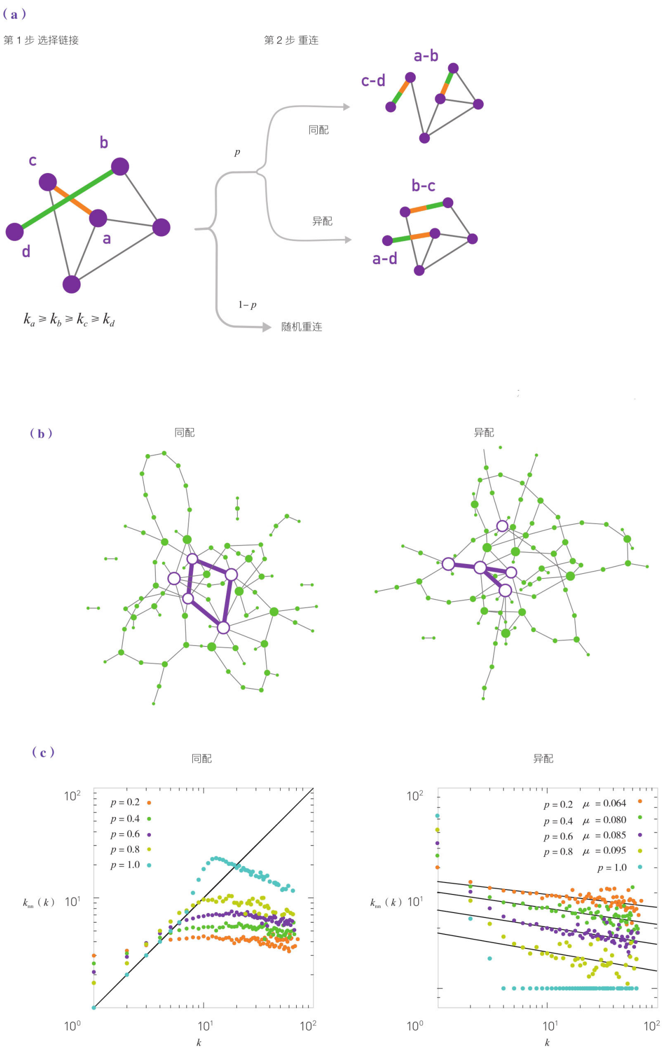


图**6-14** 调节度相关性

我们可以使用布鲁内特和索科洛夫算法来调节度相关性的大小。

（**a**）我们以概率p 执行确定性的重连，以概率$1\text{-}p$ 随机重连节点a 、b 、c 、d 。$p=1$的情况即为图**6-13** 所示，产生了最大的度相关性。$p<1$的情况引入了噪声，从而可以调节其作用的大小。

（b）$p = 0.5$所生成的典型的网络配置。

（c）一个$N = 10\ 000$、$\langle k \rangle = 1$、$\gamma = 3.0$的网络在不同p值下的$k_{nn}(k)$函数。

注意，拟合的区域不同，拟合出的相关性指数μ也不同，这一点对于同配网络更为明显。

总而言之，如果我们允许多重链接存在，静态模型（例如配置模型、隐参数模型等）就是中性的。如果我们增加生成简单网络的约束，静态模型就会产生结构性异配。要生成相关性可调节的网络，可以使用布鲁内特和索科洛夫算法等方式。本节的一个重要结论是公式**6.16** 和公式**6.18**，它们为隐参数模型和生长网络的度相关性函数提供了解析形式——二者都得出了幂律依赖于k的结论。这些结果为公式**6.10** 提供了理论支持，表明结构性和动力性因素，如初识吸引因子，都可能导致度相关性函数呈现幂律性。

6.7 度相关性的影响

如图**6-10** 所示，大部分真实网络都具有某种度相关性。社交网络是同配的，生物网络表现出结构性异配，这背后隐含了一个重要的问题：为什么我们要关心度相关性？换言之，度相关性改变了网络属性吗？何种网络属性会受到其影响？本节将讨论这些重要问题。

随机网络的一个重要属性是相变点在$\langle k \rangle = 1$时出现，从而产生了巨连通分支（**2.6**节）。图**6-15** 展示了具有不同度相关性的网络巨连通分支的相对大小，也呈现出几种不同的模式[8]，[19]，[20]：

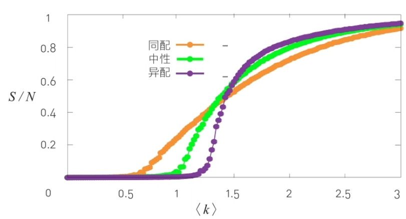


图**6-15** 度相关性与相变点

一个节点数为$N=10\ 000$的埃尔德什 - 雷尼网络的巨连通分支的相对大小（绿色曲线）。我们使用$p=0.5$的布鲁内特和索科洛夫算法对该网络重连以产生度相关性（图**6-14**）。如图所示，从同配网络过渡到异配网络，相变点被推迟了，且巨连通分支也随着k的增大而增大。每个点代表10次随机模拟的平均。

- 同配网络

同配网络的相变点在$\langle k\rangle$较小时取到，因此$\langle k\rangle<1$时就会出现巨连通分支。原因在于，如果大度节点彼此相连，则容易出现巨连通分支。

- 异配网络

异配网络的相变点推迟了，这是因为枢纽节点倾向于连接小度节点。因此，异配网络不容易产生巨连通分支。

- 巨连通分支

当$\langle k\rangle$较大时，同配网络的巨连通分支小于中性网络和异配网络的巨连通分支。事实上，同配性要求枢纽节点彼此相连，因此它们很难吸引到大量的小度节点进入巨连通分支。

巨连通分支在标度和结构上的变化会对疾病传播过程产生影响[22-24]，这一话题我们将在第**9**章讨论。事实上，如图**6-10**所示，社交网络通常是同配的。因此，大度节点产生的巨连通分支充当了疾病的“水池”，即使网络的稠密程度不足以让病毒长期存在，流行病也仍能延续。

巨连通分支的改变对网络鲁棒性亦有影响[25]。如第**7**章所讨论的，从网络中移除枢纽节点将会使网络碎片化。在同配网络中，移除枢纽节点的破坏性较小，因为很多枢纽节点共同形成一个核心节点组，从而有很多枢纽节点是冗余的。但在异配网络中，移除枢纽节点的破坏性较大，因为枢纽节点连接到大量小度节点，当枢纽节点被移除后，这些小度节点将从网络中脱落。

以下是度相关性的其他几点影响：

- **图6-16** 展示了重连后的随机网络的路径长度分布，我们对该网络重连，以展示不同度相关性的效果。如图所示，同配网络的平均路径长度短于中性网络的平均路径长度，并且网络直径d_{max}的差异更为显著：同配网络的直径明显较大。事实上，同配性倾向于让小度节点相连，于是产生了由$k=2$个节点构成的长链，从而增大了d_{max}（**图6-13c**）。

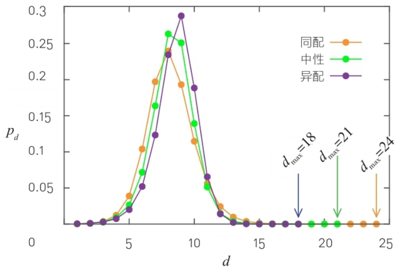


图**6-16**　度相关性和路径长度

一个$N=10\ 000$且$\langle k\rangle=3$的随机网络的路径长度分布。我们采用布鲁内特和索科洛夫算法取$p=0.5$对网络生成相关性（**图6-14**）。如图所示，当我们从异配网络转向同配网络，峰逐渐移向左侧，表明平均路径长度减短，同时直径d_{max}增长。每条曲线表示10次独立实验的均值。

- 度相关性会影响网络在遇到刺激和扰动时的稳定性[26]，以及放置在网络中的振子的同步过程[27]，[28]。
- 度相关性对于点覆盖问题（vertex cover problem）有根本性影响[29]。点覆盖问题要求寻找一个最小的节点集合，使得网络中每一条链接有至少一个端点在该集合中（**边栏6.4**），这是图论中被研究较多的问题。

边栏**6.4**

点覆盖和博物馆保安

假设有一个露天博物馆坐落在一个大型公园中，你是博物馆的馆长。你希望在路口布置保安以监控每一条路径。同时，为了节约

成本，雇用的保安越少越好。你需要多少个保安？

设N为路口的数量，$m<N$是你能在预算内雇用的保安数量。虽然有$\binom{N}{m}$种方案把m个保安部署在N个路口，但大部分方案都会留下一些监控不到的路径[31]。

若试图通过反复尝试来找到部署保安的方案，使得保安能监控所有路径，则尝试次数随着N呈指数增长。实际上，这是6个基本的NP-完全性问题之一，被称为点覆盖问题。一个网络的点覆盖是指找到网络中一个节点集合，使得每一条链接有至少一个端点在该集合中（图**6-17**）。NP-完全性的意思是目前没有任何已知算法可以比穷举（遍历每一种可能的解）更快地找到一个最小点覆盖。最小点覆盖的节点数量取决于网络拓扑，同时受到度分布和度相关性的影响[29]。

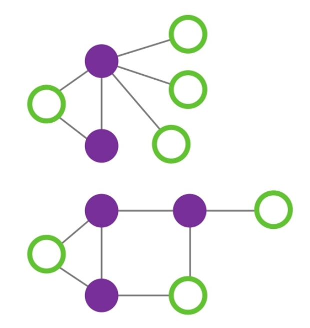


图**6-17**　最小覆盖

网络点覆盖的形式定义是，对于一个节点集合C，使得网络中每一条链接有至少一个端点在C中。一个最小点覆盖是所有满足条件的点覆盖中节点最少的点覆盖。上图中紫色节点分别展示了两个小网络中的最小点覆盖的示例。试想，如果把任意一个紫色节点变成绿色，则至少存在一条链接不连接到任何一个紫色节点。

- 度相关性影响了我们控制网络的能力，当需要实现完全控制网络时，所需的输入信号数量会改变[30]。

总而言之，度相关性不仅有学术研究价值，还影响了网络的诸多特征，并可以对网络上的很多过程产生不可忽视的影响。

6.8
小结

2001年，罗穆阿尔多·帕斯托尔 - 萨托拉斯、阿列克谢·巴斯克斯（Alexei Vazquez）和亚历山德罗·维斯皮那尼在研究互联网时首次发现度相关性 [4] ，[5] ，同时引入了度相关性函数$k_{nn}\langle k\rangle$和对$k_{nn}\langle k\rangle$的一个大小近似公式**6.10** 。一年后，金·斯奈彭（Kim Sneppen）和谢尔盖·马斯洛夫（Sergey Maslov）使用类似e_{ij} 矩阵的联合概率$p(k_i, k_j)$来描述蛋白质相互作用网络的度相关性 [32] 。2003年，马克·纽曼引入了度相关性系数 [8] ，[9] ，以及同配、中性、异配的区分。以上这些术语都源于社会科学领域 [13] ：

同配来源于选型交配（Assortative mating），即个体倾向于与类似的个体约会或结婚。例如，低收入个体与低收入个体结婚，大学生与大学生结婚。网络科学使用“同配”表达类似的意思，即节点之间基于度的相似性：在同配网络中，枢纽节点倾向于连接到其他枢纽节点，小度节点倾向于连接到其他小度节点。在网络环境中，我们也可能遇到社会科学意义下的同配性，即节点连接到具有类似属性的节点（图**6-18** ）。

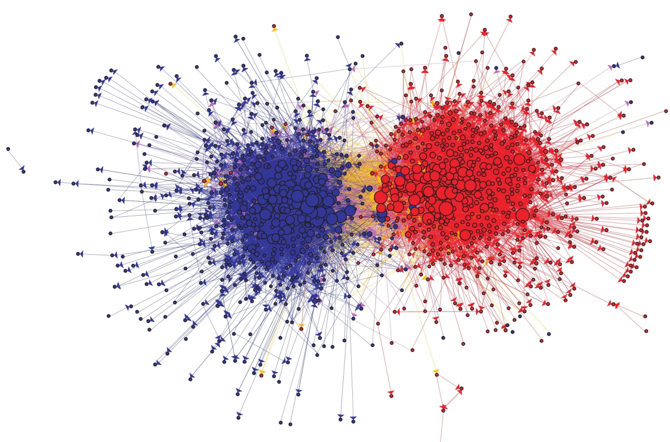


图**6-18**　政治从不是中性／中立的

美国政治博客的网络显示出社会学意义下的同配性——相似特征的节点倾向于彼此相连。图中每个蓝色节点对应自由派博客，而红色节点是保守派。自由派博客之间的链接以蓝色表示，保守派博客之间的链接以红色表示。自由派博客指向保守派博客的链接为黄色，保守派博客指向自由派博客的链接为紫色。如图所示，很少存在指向对方阵营的博客，展示出政治博客的强同配性。[33]

异配性，即个体连接到不相似的个体，在一些社会和经济系统中也很常见。性关系网络可能是最好的例子，因为大多数性关系发生在异性之间。经济系统中的交易行为通常也发生在具有不同技能的个体之间：面包师不会把面包卖给其他面包师，鞋匠很少给其他鞋匠修鞋。

综合起来，有以下三个原因使我们关心网络的度相关性（边栏**6.5**）：

边栏6.5

快速回顾：度相关性

度相关性矩阵 e_{ij}

中性网络：

$$e_{ij} = q_i q_i = \frac{k_i p_{k_i} k_j p_{k_j}}{\langle k \rangle^2}$$

度相关性函数

$$k_{nn}(k) = \sum_{k'} k' P(k' \mid k)$$

中性网络：

$$k_{nn}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

近似方式

$$k_{nn}(k) \sim k^{\mu}$$

$\mu > 0$：同配

$\mu = 0$：中性

$\mu < 0$：异配

度相关性系数

$$r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\sigma^2}$$

$r > 0$：同配

$r = 0$：中性

$r < 0$：异配

- 大多数真实网络存在度相关性（**6.5**节 ）。
- 若度相关性存在，则会改变网络的形态（**6.7**节 ）。
- 度相关性要求我们不仅要关注度分布，还要关注仅根据 p_k 无法表达的可定量分析的节点相连模式。

虽然已有很多工作尝试表征度相关性，但我们对于这一现象的理解仍未臻完善。例如，虽然**6.6**节 中介绍了一种调节度相关性的算法，但这个问题还远未得到解决。事实上，e_{ij} 矩阵包含了网络度相关性的最准确的表达，然而根据任意给定的e_{ij} 生成网络仍然很困难。

在这一章中，我们重点关注了表征相连的两个节点间相关性的$k_{nn}(k)$。实际上，一些网络中也存在高阶相关性（边栏**6.6** ），但现在，多个节点之间的相关性带来的影响还有待深入研究。

边栏**6.6**

两点相关性和三点相关性

作为网络的重要特征，完整的度相关性可以描述为这一条件概率：$P(k^{(1)}, k^{(2)}, \ldots, k^{(k)} | k)$，即一个度$k$节点连接到度$k^{(1)}$，$k^{(2)}$，…，$k^{(k)}$的其他节点的条件概率。

两点相关性

最简单的形式是本章讨论的两点相关性，即一个度k节点连接到一个度k'节点的条件概率$P(k'|k)$。对于不相关网络，这一条件概率独立于k，即$P(k'|k) = k' p_{k'} / \langle k \rangle$[18]。由于真实网络中根据数据计算$P(k'|k)$非常麻烦，更常用的做法是计算公式**6.7**定义的度相关性函数$k_{nn}(k)$。

三点相关性

三个节点的相关性由$P(k^{(1)}, k^{(2)} | k)$确定。这一条件概率与集聚系数有关。平均集聚系数$C(k)$[22]，[23]可以形式化地写成度k节点连接到度$k^{(1)}$节点和度$k^{(2)}$节点且后二者相连的概率，在所有可能的$k^{(1)}$和$k^{(2)}$取值上求平均：

$$C\langle k \rangle = \sum_{k^{(1)}, k^{(2)}} P(k^{(1)}, k^{(2)} | k) p^{k}_{k^{(1)}, k^{(2)}}$$

其中$p^{k}_{k^{(1)}, k^{(2)}}$是节点$k^{(1)}$和$k^{(2)}$在有度为k的共同邻居的前提下相连的概率[18]。中性网络的$C(k)$独立于k，即

$$C = \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{\langle k \rangle^3 N}$$

6.9 课后习题

习题1　度相关性的细致平衡

前文已将本章讨论的联合概率$e_{k,k'}$、条件概率$P(k'|k)$和概率q_k写成由以下物理量组成的表达式：节点数N、平均度$\langle k \rangle$、度k的节点数N_k，度k节点和度k'节点之间的链接数$E_{k,k'}$（注意，当$k = k'$

时，$E_{k,k'}$是链接数的2倍）。基于上述这些表达式，证明对任意网络存在

$$e_{kk'}=q_k P(k'|k)$$

习题2　星型网络

考虑一个星型网络，即1个节点连接到$N-1$个度为1的节点。假设$N \gg 1$。

（1）网络的度分布p_k是怎样的？

（2）随机选择一条链接，其存在一个端点度为k的概率q_k是多少？

（3）计算网络的度相关性系数r。计算过程可以采用习题1中的$e_{k,k'}$和$P(k'|k)$表达式。

（4）这个网络是同配的还是异配的？并解释原因。

习题3　结构截断

计算表3-1中列出的无向网络结构截断k_s。基于图6-10，判断每个网络中k_s大于还是小于最大期望度$k_{\max}$，并计算$k_{\max}$以证实这一判断。

习题4　埃尔德什-雷尼网络的度相关性

考虑第2章（边栏2.1和2.2节）介绍的随机网络的埃尔德什-雷尼模型$G(N, L)$，其中N个节点通过L条随机放置的链接相连。在该模型中，节点i和j之间存在链接的概率取决于节点l和s之间是否存在链接。

（1）写出节点i和j之间存在链接的概率e_{ij}。若已知节点l和s之间存在链接，写出节点i和j之间存在链接的条件概率。

（2）在一个小网络上，这两个概率的比值是多少？大网络呢？

（3） 如果使用埃尔德什-雷尼$G(N, p)$模型，（1）和（2）所讨论的物理量将会是怎样的？

基于（1）～（3）的结果，讨论使用$G(N, L)$模型代替$G(N, p)$模型生成节点较少的随机网络时所带来的影响。

6.10 进阶阅读6.A 度相关性系数

边栏6.2 中我们定义了度相关性系数r，作为度相关性的另一种度量方式[8]，[9]。使用单个数字来度量度相关性是很有吸引力的，因为这让不同属性和大小的网络中观测到的相关性可比。然而，为了有效使用r，我们必须了解其本质。

度相关性系数r背后的假设是，$k_{\text{nn}}(k)$可以近似表达为线性函数

$$k_{\text{nn}}(k) \sim rk \tag{6.21}$$

这不同于公式6.10 的近似方式，后者假设$k_{\text{nn}}(k)$与k成幂律关系，而且公式6.21 也存在一些问题：

- 初始吸引模型表明度相关性函数与k成幂律（公式6.18）或者对数（公式6.20）关系。隐参数模型也由公式6.16 推导出类似的幂律关系。因此，r是在对一个本质上非线性的函数强行做线性拟合，且这一线性关系假设并未得到数值模拟或解析计算的支持。事实上，如图6-19 所示，公式6.21 对同配网络和异配网络的数据拟合都不好。

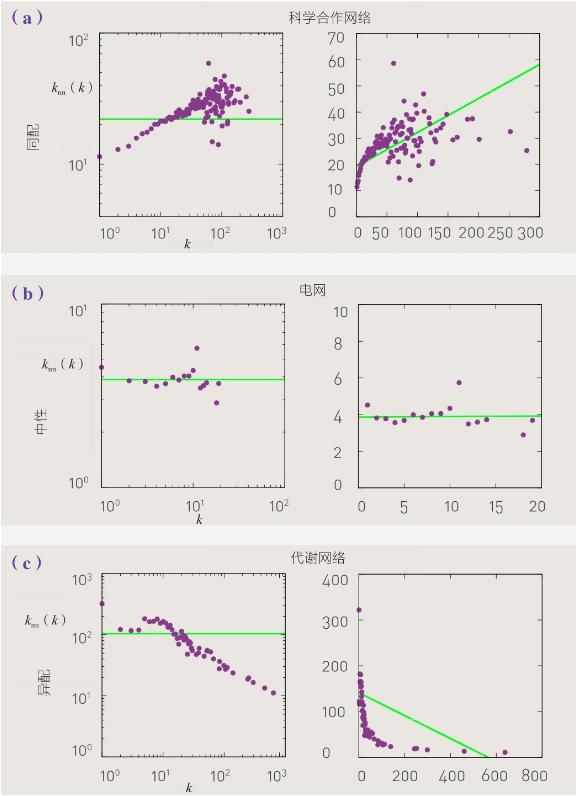


图6-19　度相关性函数

三个真实网络的度相关性函数$k_{nn}(k)$。左侧子图是双对数坐标下的$k_{nn}(k)$，以检验公式6.10的正确性。右侧子图在双线性坐标下检验公式6.21的正确性，即“$k_{nn}(k)$与k呈线性关系”这一假设的正确性。这是相关性系数r背后的假设。虚线的斜率对应相关性系数r。如右侧线性坐标图所示，公式6.21对同配网络和异配网络的数据拟合都不好。

- 如图**6-10**所示，$k_{nn}(k)$与k的关系是复杂的，常常会在k较大时出现趋势反转（因为存在结构截断），但线性拟合忽略了这种根本的复杂性。

- N较大时，最大相关模型的r归零，然而此时的网络依然存在度相关性（边栏**6.7**）。这表明度相关性系数难以度量大型网络的相关性。

边栏6.7

大型网络的问题

根据布鲁内特和索科洛夫算法可以计算无标度网络中最大（r_{max} ）和最小（r_{min} ）的相关性系数，可得[21] ：

$$r_{min} \sim \begin{cases} -c_1(\gamma, k_0) & \gamma < 2 \\ -N^{(2-\gamma)/(\gamma-1)} & 2 < \gamma < 3 \\ -N^{(\gamma-4)/(\gamma-1)} & 3 < \gamma < 4 \\ -c_2(\gamma, k_0) & 4 < \gamma, \end{cases}$$

$$r_{max} \sim \begin{cases} -N^{(-\gamma-2)/(\gamma-1)} & 2 < \gamma < \gamma_r \\ -N^{-1/(\gamma^2-1)} & \gamma_r < \gamma < 3, \end{cases}$$

其中 $\gamma_r \approx \frac{1}{2} + \sqrt{17/4} \approx 2.56$ 。

以上表达式表明：

（i） 即使该网络的重连是为了达到最大相关性，但当N 较大时，r_{min} 和r_{max} 都归零。因此，相关性系数r 不能刻画大型网络的相关性。

（ii） γ ＜2.6的无标度网络总有负值的r 。这是由结构性相关性所致（**6.4**节 ）。

鉴于r 有上述不足，我们必须分析$k_{nn}(k)$ ，以期最好地刻画大型网络的度相关性。

μ 和r 的关系

积极的一面是，r 和μ 并非相互独立。我们在10个真实网络（表**6-1** ）中计算r 和μ 以展示二者的关系，图**6-20** 是结果展示，表明μ 和r 在r 取正值时相关。然后，在r 取负值时，这一相关性消失。为了理解这一现象的本质，接下来我们将推导μ 和r 的直接关系。具体而言，我们假设公式**6.10** 是正确的，然后尝试通过相关性指数μ 确定网

络的r 的值。

表6-1　参照网络的度相关性

网络	N	r	μ
互联网	192 244	0.02	0.56
万维网	325 729	–0.05	–1.11
电网	4 941	0.003	0.0
手机通话网络	36 595	0.21	0.33
电子邮件网络	57 194	–0.08	–0.74
科学合作网络	23 133	0.13	0.16
演员网络	702 388	0.31	0.34
引文网络	449 673	–0.02	–0.18
大肠杆菌代谢网络	1 039	–0.25	–0.76
蛋白质相互作用网络	2 018	0.04	–0.1

表中列出了10个真实网络的r 和μ 的估计值。估计r 和μ 时，将有向网络都视为无向网络。另外，我们也可以对这些有向网络计算有向相关性系数（边栏**6.8** ）。

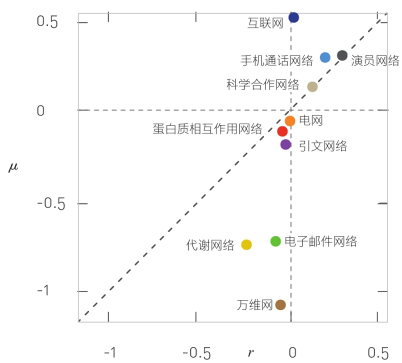


图6-20　r 与μ 之间的相关性

我们用公式**6.10** 拟合$k_{\mathrm{nn}}(k)$以估计μ ，以展示r 和μ 的关系，看看幂律是否有统计显著性。

我们利用公式6.10 确定a ，根据度分布的二阶矩

$$\langle k^2\rangle = \langle k_{\mathrm{nn}}(k)\,k\rangle = \sum_k ak^{\mu+1}p_k = a\langle k^{\mu+1}\rangle$$

可得

$$a=\frac{\langle k^2\rangle}{\langle k^{\mu+1}\rangle}$$

给定网络的μ，我们现在计算r：

$$r=\frac{\sum_k kak^{\mu}q_k-\frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2}=\frac{\sum_k ak^{\mu+2}\frac{p_k}{\langle k\rangle}-\frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2}$$

$$=\frac{\frac{\langle k^2\rangle}{\langle k^{\mu+1}\rangle}\frac{\langle k^{\mu+1}\rangle}{\langle k\rangle}-\frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2}=\frac{1}{\sigma_r^2}\frac{\langle k^2\rangle}{\langle k\rangle}\left(\frac{\langle k^{\mu+2}\rangle}{\langle k^{\mu+1}\rangle}-\frac{\langle k^2\rangle}{\langle k\rangle}\right)\quad(6.22)$$

当$\mu=0$时，最后一组括号项归零，导致$r=0$。因此，若$\mu=0$（中性网络），r也判断网络是中性的。若$k>1$，那么当$\mu>0$时，括号项取正值，则$r>0$；同样，当$\mu<0$时，括号项取负值，则$r<0$。因此r和μ对度相关性的判断类似。

总而言之，若度相关性函数服从公式**6.10**的形式，则度相关性指数μ的符号将决定系数r的符号：

$$\mu<0\to r<0$$

$$\mu=0\to r=0$$

$$\mu>0\to r>0$$

有向网络

要度量有向网络中的相关性，我们需要考虑每个节点i的特征，包括入度k_i^{in}和出度k_i^{out}。因此我们定义4个度相关性系数$r_{\text{in,in}}$、$r_{\text{in,out}}$、$r_{\text{out,in}}$和$r_{\text{out,out}}$，以刻画两个相连节点的入度和出度的所有可能的组合（图**6-21a～d**）。形式上，我们有[14]：

$$r_{\alpha,\beta}=\frac{\sum_{jk} jk\left(e_{jk}^{\alpha,\beta}-q_{\leftarrow j}^{\alpha}q_{\rightarrow k}^{\beta}\right)}{\sigma_{\leftarrow}^{\alpha}\sigma_{\rightarrow}^{\beta}} \qquad (6.23)$$

其中α 和β 可分别表示入度（in）和出度（out），$q_{\leftarrow j}^{\alpha}$ 指随机选一条链接其起点度为i 的概率，$q_{\rightarrow k}^{\beta}$ 指随机选一条链接其终点度为k的概率。$\sigma_{\leftarrow}^{\alpha}$ 和$\sigma_{\rightarrow}^{\beta}$ 是对应的标准差。公式**6.23** 的结果展示在图**6-21e** 中，给出了5个有向真实网络（表**6-1** ）的4种相关性系数。然而，如果要获得度相关性的完全表达，理想情况是同时度量4个$k_{nn}(k)$函数（边栏**6.3** ）。

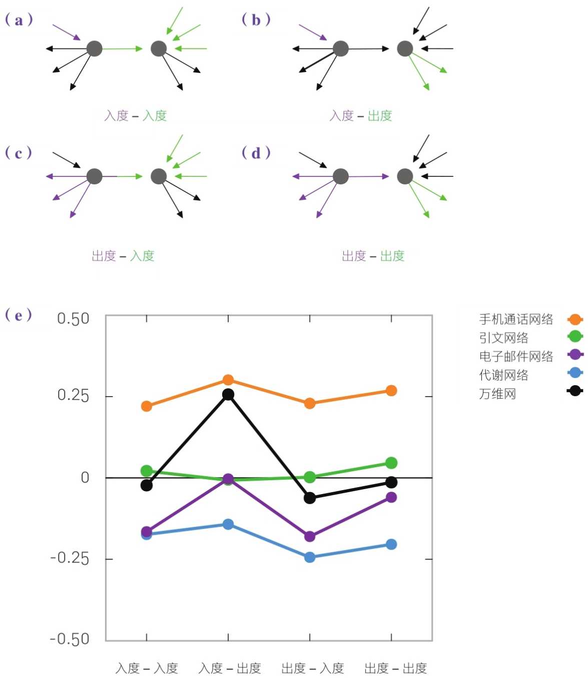


图6-21　有向相关性

（**a**）～（**d**）紫色和绿色链接表示α，β ，定义了有向网络上不同的相关性系数。

（**e**）5个有向网络的相关性。引文网络的相关性可以忽略，另外，手机通话网络的4个相关性系数表现出强同配性，代谢网络则表现出强异配性。万维网的情况比较有趣：入度－出度表现出强烈的同配倾向，而其他三种相关性系数接近0。

总而言之，度相关性系数假设$k_{nn}(k)$与k有线性关系。这一假设缺乏数值和解析支持。解析计算表明，$k_{nn}(k)$与k的关系是幂律形式（公式**6.10**）或弱的对数形式（公式**6.20**）。然而，一般来说r和μ的符号一致。因此，我们可以用r得到一个网络的潜在相关性的快速判断，然而，对底层的度相关性的准确度量依旧要求我们计算$k_{nn}(k)$。

6.11
进阶阅读6.B
结构截断

如**6.4**节所讨论的，无标度属性和度相关性的基本矛盾导致了简单网络中的结构截断。本节我们将计算结构截断如何依赖于网络规模N，并以此来推导公式**6.15**[11]。

我们首先定义

$$r_{kk'}=\frac{E_{kk'}}{m_{kk'}} \quad (6.24)$$

其中$E_{k,k'}$是度k节点和度k'节点之间链接数（$k\neq k'$）或链接数的2倍（$k=k'$），且有

$$m_{kk'}=\min\{kN_k,k'N_{k'},N_{-k}N_{-k'}\} \quad (6.25)$$

是$E_{k,k'}$可能取得的最大值。图**6-22**解释了公式**6.25**的原因。因此，我们可以将$r_{k,k'}$写作

$$r_{kk'}=\frac{E_{kk'}}{m_{kk'}}=\frac{\langle k\rangle e_{kk'}}{\min\{kP(k),k'P(k'),NP(k)P(k')\}} \quad (6.26)$$

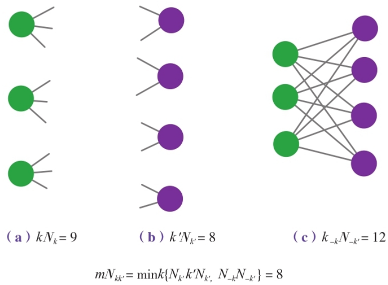


图6-22　计算$m_{k,k'}$

两组节点之间可能的最大链接数。图中展示了两组节点，度分别为$k=3$和$k'=2$。两组节点之间的链接总数不能超过：

（**a**）$k=3$组可用的链接总数，即$kN_k=9$。

（**b**）$k'=2$组可用的链接总数，即$k'N_{k'}=8$。

（**c**）两组节点之间可放置的链接总数，即$N_k N_{k'}$。

这三者的最小值是（b）中的$k'N_{k'}=8$。

因为$m_{k,k'}$是$E_{k,k'}$可能取得的最大值，对任意k和k'必有$r_{k,k'}\leq 1$。严格来说，简单网络中任意k和k'取值不可能有$r_{k,k'}>1$。然而，对于一些网络的一些k和k'取值，$r_{k,k'}$可以大于1。这显然是不合理的，说明网络配置存在矛盾。因此，我们将结构截断k_s定义为下式的解：

$$r_{k_s k_s}=1 \tag{6.27}$$

注意，只要$k>Np_{k'}$且$k'>Np_k$，这一限制就会起作用，从而使$r_{k,k'}$的表达式变成

$$r_{kk'}=\frac{\langle k\rangle e_{kk'}}{Np_k p_{k'}} \tag{6.28}$$

无标度网络中满足条件的取值区间是$k, k'>(aN)^{1/(\gamma+1)}$，其中$a$是关于$p_k$的常量。这个值低于自然截断，因此，该式提供了结构截断的下界。也就是说，任何时候只要度分布的截断低于这一限制，则一定满足条件$r_{k,k'}<1$。

中性网络的联合分布可以因子化为

$$e_{kk'} = \frac{kk' p_k p_{k'}}{\langle k \rangle^2} \quad (6.29)$$

则公式6.28 中的比例变成

$$r_{kk'} = \frac{kk'}{\langle k \rangle N} \quad (6.30)$$

因此，满足条件$r_{k,k}{'} \leq 1$的结构截断应有以下形式 [11] ，[34] ，[35] ，[36] ：

$$k_s(N) \sim (\langle k \rangle N)^{1/2} \quad (6.31)$$

这正是公式6.15 。另外公式6.31 与底层网络的度分布无关。因此，对于无标度网络，k_s (N)独立于度指r 数。

6.12 参考文献

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, RS Judson,JR Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart,A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G.Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg.A comprehensive analysis of protein- protein interactions in Saccharomyces cerevisiae. *Nature* , 403: 623-627, 2000.

[2] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M.Marcotte, D. Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Res* ., 28: 289-29, 2000.

[3] H. Jeong, S.P. Mason, A.-L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature* , 411: 41-42, 2001.

[4] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett*

., 87: 258701, 2001.

[5] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of Internet. *Phys. Rev. E,* 65: 066130, 2002.

[6] S.L. Feld. Why your friends have more friends than you do. *American Journal of Sociology* , 96: 1464–1477, 1991.

[7] E.W. Zuckerman and J.T. Jost. What makes you think you're so popular? Self evaluation maintenance and the subjective side of the "friendship paradox". *Social Psychology Quarterly* , 64: 207-223, 2001.

[8] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett* ., 89: 208701, 2002.

[9] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev* . *E* , 67: 026126, 2003.

[10] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological pattern in complex networks: Correlation profile of the Internet. *Physica A* , 333: 529-540, 2004.

[11] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Cut-offs and finite size effects in scale-free networks. *Eur. Phys. J. B* , 38: 205, 2004.

[12] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E* , 68:036122, 2003.

[13] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: homophily in social networks. *Annual Review of Sociology* , 27:415-444, 2001.

[14] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski. Edge direction and the structure of networks. *PNAS* , 107: 10815, 2010.

[15] A. Barrat and R. Pastor-Satorras. Rate equation approach for correlations in growing network models. *Phys. Rev.E* , 71: 036127, 2005.

[16] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Adv. Phys* ., 51: 1079, 2002.

[17] J. Berg and M. Lässig. Correlated random networks. *Phys. Rev. Lett* ., 89: 228701, 2002.

[18] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Phys. Rev. E* , 68: 036112, 2003.

[19] R. Xulvi-Brunet and I. M. Sokolov. Reshuffling scale-free networks: From random to assortative. *Phys. Rev. E* , 70: 066102, 2004.

[20] R. Xulvi-Brunet and I. M. Sokolov. Changing correlations in networks: assortativity and dissortativity. *Acta Phys. Pol. B* , 36: 1431, 2005.

[21] J. Menche, A. Valleriani, and R. Lipowsky. Asymptotic properties of degree-correlated scale-free networks. *Phys. Rev. E* , 81: 046103, 2010.

[22] V. M. Eguíluz and K. Klemm. Epidemic threshold in structured scale-free networks. *Phys. Rev. Lett* ., 89:108701, 2002.

[23] M. Boguñá and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E* , 66: 047104, 2002.

[24] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett* ., 90: 028701, 2003.

[25] A. Vázquez and Y. Moreno. Resilience to damage of graphs with degree correlations. *Phys. Rev. E* , 67: 015101R, 2003.

[26] S.J. Wang, A.C. Wu, Z.X. Wu, X.J. Xu, and Y.H. Wang. Response of degree-correlated scale-free networks to stimuli. *Phys. Rev. E* , 75: 046113, 2007.

[27] F. Sorrentino, M. Di Bernardo, G. Cuellar, and S.Boccaletti. Synchronization in weighted scale-free networks with degree-degree correlation. *Physica D* , 224: 123, 2006.

[28] M. Di Bernardo, F. Garofalo, and F. Sorrentino. Effects of degree correlation on the synchronization of networks of oscillators. *Int. J. Bifurcation Chaos Appl. Sci. Eng* ., 17: 3499, 2007.

[29] A. Vazquez and M. Weigt. Computational complexity arising from degree correlations in networks. *Phys. Rev. E* , 67:027101, 2003.

[30] M. Posfai, Y Y. Liu, J-J Slotine, and A.-L. Barabási. Effect of correlations on network controllability. *Scientific Reports* , 3:1067, 2013.

[31] M. Weigt and A. K. Hartmann. The number of guards needed by a museum: A phase transition in vertex covering of random graphs. *Phys. Rev. Lett* ., 84: 6118, 2000.

[32] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science* , 296: 910-913, 2002.

[33] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog (2005) .

[34] J. Park and M. E. J. Newman. The origin of degree correlations in the Internet and other networks. *Phys. Rev. E* , 66:026112, 2003.

[35] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* , 6: 125, 2002.

[36] Z. Burda and Z. Krzywicki. Uncorrelated random

networks. *Phys. Rev. E* , 67: 046118, 2003.

第7章
网络鲁棒性



图7-0　网络与艺术：**Facebook**用户

本图由多伦多数据科学家保罗·巴特勒（Paul　Butler）在Facebook实习期间制作，描述了连接用户的网络。它突出了大陆内部及不同大陆之间的链接。美国、欧洲和印度很明显有着更密集的内部链接，同样明显的还有那些缺失链接的区域，例如缺少互联网接入的非洲。

本章索引

7.1　本章概览 / 337

7.2　渗流理论 / 339

7.3　无标度网络的鲁棒性 / 345

7.4　攻击耐受性 / 351

7.5　级联故障 / 355

7.6　对级联故障建模 / 360

7.7　构筑鲁棒性 / 365

7.8　小结：阿喀琉斯之踵 / 370

7.9　课后习题 / 373

7.10　进阶阅读7.A
无标度网络中的渗流 / 376

7.11　进阶阅读7.B
莫洛伊 - 里德准则 / 378

7.12　进阶阅读7.C
随机故障下的临界阈值 / 380

7.13　进阶阅读7.D
分解有限规模的无标度网络 / 383

7.14　进阶阅读7.E
真实网络对攻击和故障的耐受性 / 385

7.15　进阶阅读7.F
攻击阈值 / 387

7.16　进阶阅读7.G
最优度分布 / 390

7.17　参考文献 / 392

本章致谢

马顿·波什福伊　**MÁRTON PÓSFAI**

萨拉·莫里森　**SARAH MORRISON**

加布里埃莱·穆塞拉　**GABRIELE MUSELLA**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

妮科尔·萨迈　**NICOLE SAMAY**

菲利普·赫费尔　**PHILIPP HOEVEL**

罗伯塔·西纳特拉　**ROBERTA SINATRA**

7.1 本章概览

组件故障或者失效都可以破坏人类设计的系统。比如，汽车引擎的一个部件失灵可能会使车在半路抛锚，你不得不打电话叫拖车来帮忙；又如，计算机芯片中的布线错误会导致计算机完全不能工作。然而，很多自然和社会系统有一种出色的能力：即使它们的一些组成部分失效了，它们仍然能够维持基本功能。事实上，我们的细胞中有无数蛋白质非正常折叠和化学反应缺失这样的错误，但我们很少注意到这些错误导致的后果。与之类似，大型机构在很多员工缺席的时候也仍然可以运转。理解这种鲁棒性的起源对许多学科都非常重要：

- 鲁棒性是生物学和药学的核心问题。它可以帮助我们理解，为什么有些基因突变会引发疾病，而另一些却不能。
- 社会科学家和经济学家都很关心鲁棒性，因为它关系到人类社会和组织在经历饥荒、战争、社会经济秩序变更时的稳定性。
- 鲁棒性是生态学家和环境科学家面临的一个核心问题，据此，他们希望可以预测人类活动对生态系统的破坏何时会造成系统性失效。
- 鲁棒性是工程学的最终目标。工程师期望设计出的通信系统、汽车或者飞机在偶尔遇到组件故障时，仍然可以实现其基本功能。



图7-1　复杂网络中的阿喀琉斯之踵

此图为《自然》杂志2010年7月27日的封面，这本杂志中发表了一篇论文——《复杂网络的攻击与容错性》（*Error and Attack Tolerance of Complex Networks*），该论文开启了对网络的鲁棒性的科学探索[1]。

网络在生物体系、社会制度和技术系统的鲁棒性中起着关键性的作用。事实上，细胞的鲁棒性内嵌在一个复杂的、包含了调节机制、信息传导和新陈代谢的网络里；社会体系的自我复原力也和它背后错综复杂的，包含了社交关系、职业关系和通信关系的网络密不可分；如果不对那维系着每一个物种的食物链进行仔细分析，我们也无法理解生态系统作为一个整体的生命力。在大自然寻求鲁棒性的时候，它总会诉诸网络。

本章的目的是了解网络在维持复杂系统的鲁棒性中起到的作用。我们会向读者展示，当系统遇到随机故障或蓄意攻击的时候，系统的底层网络结构会对其耐受能力起到至关重要的作用。我们将探究网络在级联故障中所产生的作用，这种级联故障经常对真实系统造成破坏。最重要的是，我们将揭示在复杂网络里，容错性、耐攻击性和级联故障的出现所遵循的规律是一致的。理解这些规律有助于我们了解广泛的复杂系统的鲁棒性。



图7-2　鲁棒的，鲁棒性

鲁棒的英文robust源于拉丁语Quercus　Robur，也就是橡树，古时候，橡树代表着力量与长寿。图中的树长在匈牙利的迪欧什淮斯洛（Diosviszlo）村附近，被记录在一个收录了匈牙利境内最古老、最大的树木的网站上。感谢吉奥尔吉·波什福伊（György Pósfai）提供图片。

7.2
渗流理论

移除单一节点对网络完整性的影响是有限的（图7-3a　）。然而，如果移除网络中的多个节点，就可以把一个网络拆分成多个独立的连通分支（图7-3d　　）。显然，移除更多的节点，我们就更有可能破坏网络。于是，我们不禁要问：要移除多少个节点，我们才能把一个网络分解成一些相互独立的连通分支呢？举一个具体的例子：在互联网中，要破坏多少路由器，我们才能把互联网分解成相互不能沟通的计算机集群（簇）呢？若想回答这些问题，我们必须先熟悉研究网络鲁棒性的数学基础——渗流理论（percolation theory）。

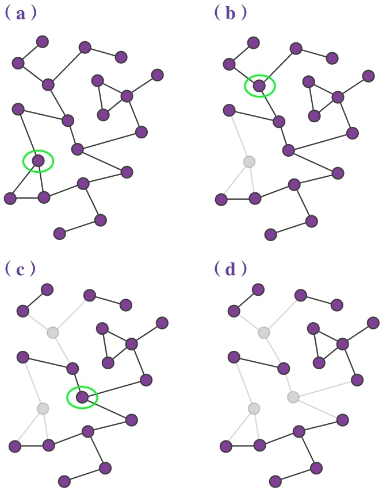


图7-3　节点移除的影响

此图显示的是，在一个小网络中，通过移除节点使网络逐渐碎片化的过程。在每一个子图中，我们移除一个不同的节点（用绿色圆圈表示）和它所有的链接。虽然移除第一个节点只对网络完整性产生有限的影响，但移除第二个节点就让2个小簇脱离网络其余部分。当第三个节点被移除时，网络被分解为5个不连通的簇，大小分别是$S=2,2,2,5,6$。

渗流

渗流理论是统计物理学和数学领域衍生出的一个成熟的分支领域[2-5]。图7-4a、b 展示了一个典型的渗流理论可以解决的问题。在方形网格的每个交叉点上，我们以一定的概率p 来放置石子。我们认为相邻的石子相互连接，两个或两个以上相互连接的石子形成簇（cluster）。鉴于每个石子放置的位置是随机的，我们提出以下问题：

- 规模最大的簇的期望值是多少？
- 所有簇的平均值是多少？

很显然，概率p 越高，簇越大。渗流理论的一个关键推断是，簇的规模并不随着概率p 的升高而逐渐增加。相反，存在一个较宽的概

率p 区间，网格里只有大量的小簇（图**7-4a** ）。当p 升高到一个临界值p_c 时，这些小簇会增长并且凝聚，促使产生一个大簇。当这个大簇扩张到网格的边缘时，我们称之为渗流簇（percolating cluster）。换言之，在临界值p_c ，我们观察到相变——许多小簇凝聚成一个可以渗透整个网格的渗流簇（图**7-4b** ）。

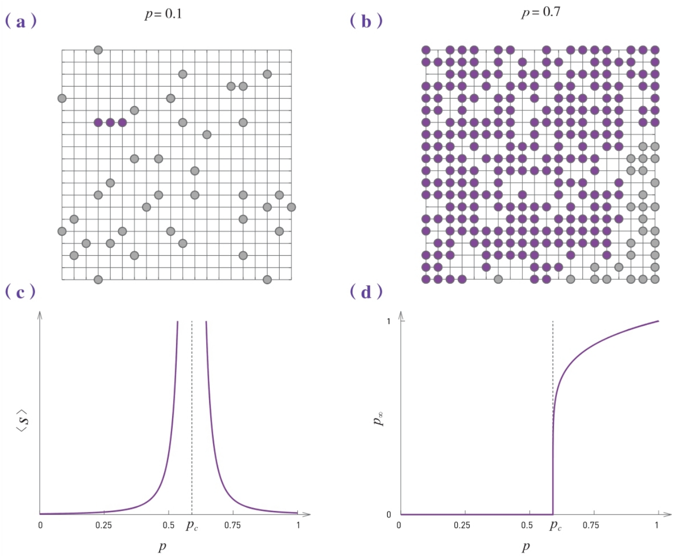


图**7-4**　渗流

渗流理论里一个经典的问题——在方形网格上以概率p 随机放置石子。

（**a**） 当概率p 较小时，大部分石子是孤立的。在这种情况下，最大的簇只有用紫色标识的那3个节点。

（**b**） 当概率p 较大时，大部分（但不是全部）石子属于一个单一的簇，仍用紫色标识。这个簇被称为渗流簇，它贯穿了整个网格（参见图**7-6** ）。

（**c**） 簇的平均规模$\langle s \rangle$ 与概率p 的函数。当我们从左侧逼近p_c 时，很多小的簇凝聚在一起，$\langle s \rangle$ 发散，正如公式**7.1** 所示。当从右侧逼近p_c 时，也有同样的情况，只是我们计算$\langle s \rangle$ 时应不考虑渗流簇。在临界点p_c 两侧的发散可以用相同的指数γ_p 来描述。

（**d**） 一个石子属于最大连通分支的概率P_∞ 对p 的依赖关系。当$p < p_c$ 时，所有的连通分支都很小，所以P_∞ 是零。一旦p 达到p_c 时，就会出现一个巨连通分支。因此，在$P > p_c$ 的情况下，如公式**7.2** 示，存在一个非零的概率使一个节点属于最大连通分支。

为了量化这个相变的性质，我们专注于三个量。

- 簇的规模的平均值：$\langle s\rangle$

根据渗流理论，全体有限规模的簇的规模的平均值遵循：

$$\langle s\rangle \sim |p-p_c|^{-\gamma_p} \qquad (7.1)$$

换句话说，在接近p_c时，簇的规模的平均值是发散的（图**7-4c**）。

- 序参量：P_∞

随机选择一个石子，序参量P_∞是这个石子属于最大簇的概率，它遵循：

$$P_\infty \sim (p-p_c)^{\beta_p} \qquad (7.2)$$

因此，当p逐渐减小到p_c时，一个石子属于最大簇的概率降为零（图**7-4d**）。

- 关联长度：ξ

属于同一簇的两颗石子之间的平均距离，它遵循：

$$\xi \sim |p-p_c|^{-\nu} \qquad (7.3)$$

因此，当$p<p_c$时，同一簇石子之间的距离是有限的。当$p=p_c$时，这个距离是发散的。这意味着，在临界点p_c，最大簇的规模是无穷大的，因而可以渗流整个网格。

指数γ_p、β_p和ν刻画了系统在临界点p_c附近的行为，因此它们被称为临界指数。渗流理论预测，这些指数是普适的，不依赖于网格的性质或具体的p_c值。因此，无论我们把石子放在三角形还是六角形的网格上，刻画$\langle s\rangle$、P_∞和ξ这些量的γ_p、β_p和ν指数都是相同的。

为了更好地理解这种普适性，考虑下面的一些例子：

- 临界值p_c 依赖于网格类型，因此不具有普适性。例如，对于一个二维方形网格（图**7-4**）来说，$p_c \approx 0.593$；而对于二维三角网格来说，$p_c = 1/2$（点渗流）。
- 网格的维度也会改变临界值p_c：在二维（$d = 2$）方形网格中，$p_c \approx 0.593$；而在三维（$d = 3$）立方网格中，$p_c \approx 0.3116$。因此，对于三维网格来说，我们只需要在较少的交叉点上放置石子就可以达到渗流相变。
- 和临界值p_c 不同，临界指数不依赖于网格类型，只依赖于网格的维度。在如图**7-4** 所示的二维网格中，我们有$\gamma_p = 43/18$、$\beta_p = 5/36$和$\nu = 4/3$。在三维网格中，我们有$\gamma_p = 1.80$、$\beta_p = 0.41$和$\nu = 0.88$。对于任意$d > 6$的网格，我们有$\gamma_p = 1$、$\beta_p = 1$和$\nu = 1/2$。因此，当维度足够高时，临界指数和维度无关[2]。

逆渗流相变与鲁棒性

研究鲁棒性的人所关注的一个有趣的现象是，节点故障对网络完整性的影响。我们可以用渗流理论来描述这个过程。

我们把方形网格视为一个网络，网格中的交叉点即为网络的节点（图**7-5**）。我们随机移除占比为f 的节点，然后观察它们的移除如何影响网格的完整性。

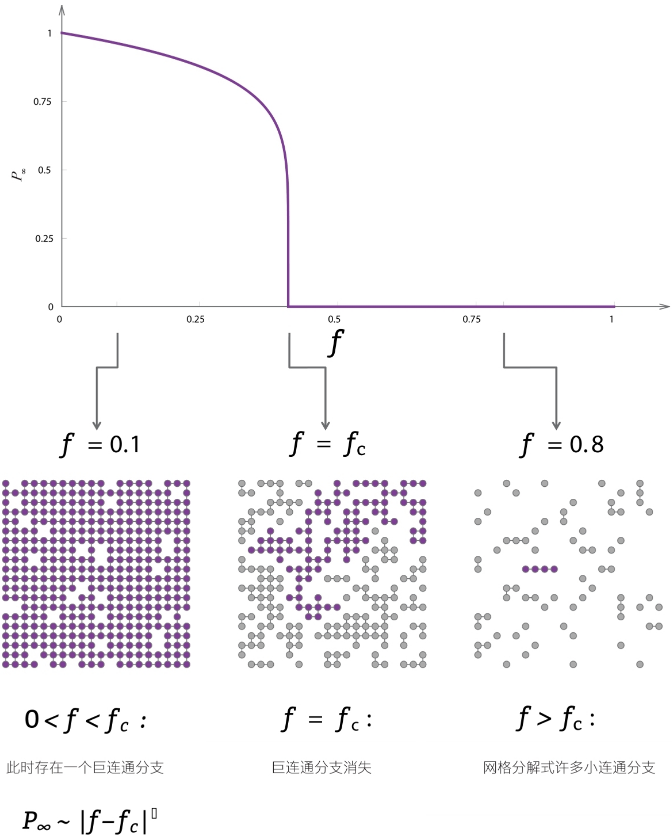


图7-5　作为逆渗流过程的网络分解

节点移除带来的影响可以用图7-4 所讨论的渗流过程的逆过程来解释。我们从一个方形网格开始，把网格的交叉点当作网络的节点，随机地选取并移除占比为f 的节点，然后测量由剩余节点组成的巨大连通分支的规模。这个规模可以被概率P_{∞} 刻画，这一概率即是随机选取一个节点，这个节点属于巨大连通分支的概率。底部的子图显示了我们所观测到的网络。在每一个子图下方标注了网络所处的相。

如果f 较小，那么缺少的节点对网络的损害不大。然而，当f 增大时，很多节点块从巨连通分支中脱离出来。最后，若f 足够大，巨连通分支会分解成很多不相连的小连通分支（图7-5 ）。

这种碎片化的过程并不是渐进的，而是存在临界阈值：对于任意$f < f_c$ ，存在一个巨连通分支。但一旦f超过了f_c ，巨连通分支消失。序参量P_{∞} （一个节点属于巨连通分支的概率）对f的依赖关系说明了这

一点（图**7-5**）：当$f<f_c$时，$P_\infty>0$；但当$f\geq f_c$时，$P_\infty=0$。这里，临界指数γ_p、β_p和ν与公式**8.1**～**8.3**中的一致。事实上，只要令$f=1-p$，这两个过程就可以相互映射。

那么，如果我们的网络并不像方形网格这样规则，会发生什么呢？正如我们将在后续章节中看到的，这个问题的答案依赖于具体的网络拓扑结构。然而，对随机网络来说，渗流理论依然可以提供答案：在随机移除节点的情况下，随机网络的标度指数与无限维网格上的渗流的标度指数相同。因此，随机网络的临界指数是：$\gamma_p=1$、$\beta_p=1$和$\nu=1/2$，对应前文提到的$d>6$的渗流。对于无标度网络的临界指数，请见进阶阅读**7.A**。

综上所述，随机节点的移除对网络造成的破坏并不是一个渐进的过程。相反，移除一小部分节点对网络完整性的影响非常有限。但一旦被移除节点的比例达到一个临界阈值，网络就会迅速地分解为不相连的连通分支。换言之，随机的节点移除引发了网络从连通到碎片的相变。在规则网络和随机网络中，我们都可以使用渗流理论的工具来刻画这种相变。然而，对无标度网络来说，上述现象的核心部分发生了变化。我们将在下一节进行讨论。

边栏**7.1**

从森林火灾到渗流理论

我们可以使用火在森林中的传播来说明渗流理论的基本概念。假设图**7-4a**、**b**中的每一颗石子都是一棵树，网格就是森林。如果一棵树着火，就会点燃附近的树木，而这些新着火的树又会点燃它们的“邻居”。大火蔓延，直到每一棵树的“邻居”都着火。因此我们要问：如果点燃一棵树，森林被烧毁的比例是多少？以及，火需要多长时间才能熄灭？

答案取决于树的密度，由参数p控制。在p较小的情况下，森林由许多分散的小片树林组成（$p=0.55$，图**7-6a**）。因此，点燃任何一颗树最多烧毁一小片树林，火也会很快熄灭。在p较大的情

况，大多数树都属于同一个大簇，因此火灾会迅速席卷稠密的森林（p =0.62，图**7-6c**）。

模拟结果表明，存在一个关键的阈值p_c。当处于这个阈值时，火灾需要很长的时间才能结束。这个阈值是渗流问题的核心。事实上，在$p = p_c$ 时，许多小簇开始凝聚，巨连通分支初现（图**7-6b**）。因此，火必须沿着一条蜿蜒绵长的路径，才能到达这个松散的大簇上的每一颗树，这个过程相当耗时。

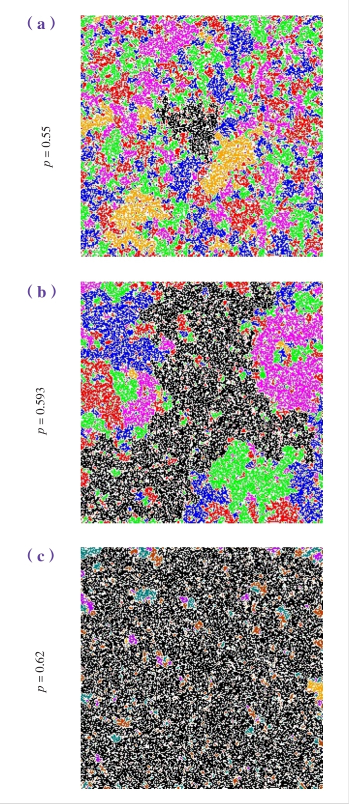


图**7-6** 森林火灾

在改变代表树的密度的参数p 后，网络中出现巨连通分支。该网格有250×250个格点，每一个子图都对应着p_c 邻域附近的一个不同的p 值。网络中的最大连通分支用黑色表示。当$p < p_c$ 时，最大连通分支非常小，如图a所示。如果网格代表森林，石子代表森林中的树，那么森林中任意一处着火都只能烧毁很小一部分树木，并且会很快熄灭。如图b所示当$p_c \approx 0.593$时，最大的簇贯穿了整个网格，所以火会

逐渐地燃烧至整个森林。当p增大到高于p_c时，很多石子（树木）被连接到最大连通分支上来，如图c所示的$p=0.62$的情况。此时，火可以迅速烧毁大部分森林，并很快地熄灭。

7.3 无标度网络的鲁棒性

渗流理论主要关注节点度相同的规则网格，或者节点度接近的随机网络。然而，无标度网络是怎样一种情况呢？网络中的枢纽节点将如何影响渗流相变呢？

为了回答这些问题，让我们来考虑路由器层面的互联网，随机选取一些路由器节点并逐个去掉。渗流理论告诉我们，一旦被移除的节点的数量达到一个阈值f_c时，互联网会分解成许多不相连的子网络（图**7-5**）。然而，数值模拟的结果却相反：即便在很多节点被移除的情况下，互联网并没有被分解。巨连通分支的规模随着节点的移除逐渐变小，最终在$f=1$时消失（图**7-5a**）。这意味着，互联网的底层网络结构对随机的节点移除有着不寻常的鲁棒性：必须移除所有的节点才能摧毁巨连通分支。这个结论不同于网格上的渗流过程——网格上的渗流理论表明，在移除一定的节点后，网络就会被分解。

以上观察到的现象并不仅仅存在于互联网中。我们在一个度指数$\gamma=2.5$的无标度网络上重复了上面的测试，结果观察到了完全一样的模式（图**7-7b**）：在随机移除节点时，巨连通分支并没有在小于1的某个f_c值处分解，而是在接近$f=1$时才逐渐消失（在线资源**7.1**）。这意味着互联网所体现的鲁棒性来源于它的无标度拓扑结构。这一节的目的就是要揭示和量化这不同寻常的鲁棒性的起源。

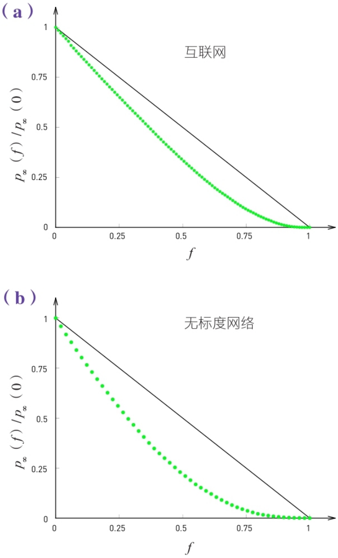
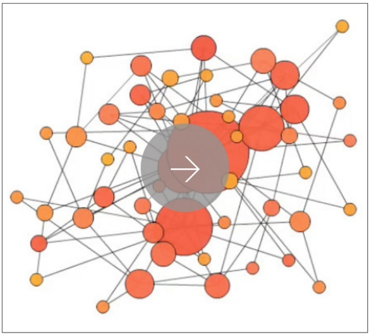


图7-7　无标度网络的鲁棒性

（a）　此图描述的是，当占比为f　的路由器被随机移除后，属于互联网巨连通分支的路由器所占的比例。比率$P_\infty(f)/P_\infty(0)$给出了巨连通分支的相对大小。模拟中使用的是路由器层面的互联网拓扑结构，如表**3-1** 所示。

（b）　此图描述的是，在一个无标度网络中，$\gamma=2.5$，$N=10\ 000$，并且$k_{min}=1$，当网络中占比为f　的节点被随机移除后，属于巨连通分支的剩余节点所占的比例。

这些图表明，互联网（或者无标度网络）在移除一定比例的节点后，网络并不会被分解，需要移除几乎所有的节点（即$f_c=1$）才能分解这些网络。



在线资源**7.1**　在节点故障情况下的无标度网络

为了说明了无标度网络的鲁棒性，我们可以从在线资源**3.1** 中构建的网络开始，即由巴拉巴西 - 阿尔伯特模型生成的无标度网络开始。在这个网络里，我们随机选择节点并逐一移除。正如短片表现的，即使移除了相当大一部分的节点，网络也并没有被裂解。感谢王大顺提供可视化效果。

扫描二维码下载“湛庐阅读”App，回复书名，深度了解无标度网络。

莫洛伊 - 里德准则（Molloy-Reed Criterion）

若想理解互联网和无标度网络为什么会有异常高的阈值f_c，我们就需要计算有着任意度分布的网络中的f_c。我们基于一个简单的观察：如果一个网络中存在巨连通分支，那么在这个巨连通分支中的大部分节点必须至少有两个邻居（图**7-8**）。这就是莫洛伊 - 里德准则（进阶阅读**7.B**）：如果满足上述条件，在一个随机网络中就会存在巨连通分支[6]：



图**7-8** 莫洛伊 - 里德准则

每个人都必须握着另外两个人的手才能形成一条链。与之类似，在网络中要存在一个巨连通分支，平均来说，每一个节点应该至少有两个邻居。莫洛伊 - 里德准则（公式**7.4**）利用了这个属性，使我们可以计算网络分解时的临界点。具体推导请见进阶阅读**7.B**。

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} > 2 \qquad (7.4)$$

在$\kappa < 2$的网络中不存在巨连通分支，而是由很多不相连的连通分支组成。莫洛伊－里德准则（公式**7.4**）以巨连通分支存在与否来表征网络的完整性，从而把网络的完整性与$\langle k\rangle$和$\langle k^2\rangle$联系起来。这个准则对任何度分布p_k都是有效的。

为了验证公式**7.4**的有效性，我们考虑随机网络的情况。因为$\langle k^2\rangle = \langle k\rangle(\langle k\rangle + 1)$，于是一个随机网络中存在巨连通分支，只要满足：

$$\kappa = \frac{\langle k^2\rangle}{\langle k\rangle} = \frac{\langle k\rangle(\langle k\rangle + 1)}{\langle k\rangle} = \langle k\rangle + 1 > 2 \tag{7.5}$$

换而言之，

$$\langle k\rangle > 1 \tag{7.6}$$

这一断言和存在巨连通分支的必要条件（公式**2.10**）一致。

临界阈值

为了理解如图**7-7**所示的鲁棒性的数学基础，我们需要寻找无标度网络在哪个阈值时失去巨连通分支。在一个有任意度分布的网络上使用莫洛伊－里德准则，我们发现临界阈值符合[7]（进阶阅读**7.C**）：

$$f_c = 1 - \frac{1}{\frac{\langle k^2\rangle}{\langle k\rangle} - 1} \tag{7.7}$$

从公式**7.7**中得到的最重要的信息是：临界阈值f_c只取决于$\langle k\rangle$和$\langle k^2\rangle$，这两个量只由度分布p_k决定。

我们用公式**7.7**来计算随机网络分解的阈值。利用$\langle k^2\rangle = \langle k\rangle(\langle k\rangle + 1)$，我们可以得到（进阶阅读**7.D**）：

$$f_c^{\mathrm{ER}} = 1 - \frac{1}{\langle k\rangle} \tag{7.8}$$

因此，随机网络越稠密，阈值f_c越高，也就需要移除更多的节点

才能裂解网络。更进一步，公式**7.8** 表明f_c 总是介于0和1之间。因此，随机网络在移除一定比例的节点后总是可以被分解的。

公式**7.7** 可以帮助我们理解图**7-7** 所展现的强鲁棒性的原因。事实上，对于γ <3的无标度网络来说，它的二阶矩$\langle k^2\rangle$ 在$N\to\infty$的时候是发散的。把$\langle k^2\rangle\to\infty$代入公式**7.7** ，我们发现$f_c$ 收敛到$f_c=1$。这意味着，要想分解一个无标度网络，我们必须移除它所有的节点。换言之，随机移除一部分节点无法分解一个无标度网络。

为了更好地理解这个结果，我们将$\langle k\rangle$ 和$\langle k^2\rangle$ 换成无标度网络的一些参数：度指数γ 、最小的度$k_{\min}$ 、最大的度$k_{\max}$ 。我们得到：

$$f_c=\begin{cases}1-\dfrac{1}{\dfrac{\gamma-2}{3-\gamma}k_{\min}^{\gamma-2}k_{\max}^{3-\gamma}-1} & 2<\gamma<3\\[2ex] 1-\dfrac{1}{\dfrac{\gamma-2}{3-\gamma}k_{\min}-1} & \gamma>3\end{cases}\qquad(7.9)$$

根据公式**7.9** 可以得到如下结果（图**7-9** ）：

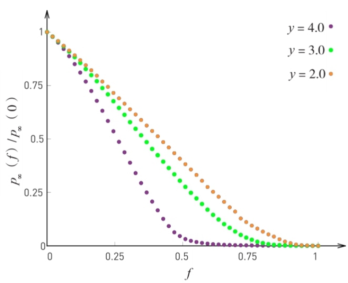


图**7-9** 鲁棒性和度指数

此图描述的是，在度指数为γ 的无标度网络中移除占比为f 的节点后，网络中剩余的某个节点属于巨连通分支的概率。当γ =4时，我们观测到的临界点$f_c\approx 2/3$，如公式**7.9** 所示。然而，当γ <3时，我们有$f_c\to 1$。网络是由配置模型生成的，其中$k_{\min}=2$，$N=10\ 000$。

- 当γ >3时，临界阈值f_c 只取决于γ 和$k_{\min}$ 。因此，f_c 与网络的规模N 无关。在这种情况下，无标度网络和随机网络的行为类似，当一部分节点被移除时会分解。

- 当$\gamma < 3$时，根据公式**3.18**，k_{max}在N较大时会发散。因此，在$N \to \infty$的极限下，公式**7.9**表明$f_c \to 1$。换言之，要想分解一个无限大的无标度网络，我们必须移除所有的节点。

公式**7.6**~**7.9**是本章的主要结果，指出无标度网络可以承受任意程度的随机节点移除而不被分解。枢纽节点的存在致使了这一不同寻常的鲁棒性。事实上，正如定义一样，随机节点移除与节点的度无关，移除大度节点和小度节点的概率相同。但是，无标度网络中的小度节点远远多于枢纽节点。因此，随机移除节点往往会移除众多小度节点中的一个，因为随机选择到少数几个枢纽节点之一的概率几乎可以忽略不计。这些小度节点对网络的完整性几乎没有贡献，因此去除这些节点并不会造成巨大的破坏性。

让我们回到图**3-6**中的机场比喻，如果随机选择一个机场，然后关闭它，我们关闭的很可能是众多小机场中的某一个。世界上的其他地方很难注意到这个小机场的消失：你仍然可以从纽约飞到东京，或从洛杉矶飞到里约热内卢。

有限网络的鲁棒性

公式**7.9**告诉我们，无标度网络在$k_{max} \to \infty$时（也就是在$N \to \infty$的极限下），f_c收敛到1。虽然我们关心的很多实际网络都非常大，但它们的规模仍然是有限的。那么我们所观察到的异常的鲁棒性对有限大的网络有什么意义呢？为了回答这个问题，我们把公式**3.18**代入公式**7.9**，可以得到f_c对网络规模N的依赖关系（进阶阅读**7.C**）：

$$f_c = 1 - \frac{C}{N^{\frac{3-\gamma}{\gamma-1}}} \qquad (7.10)$$

其中常数C包含了所有不依赖于N的项。公式**7.10**指出，网络越大，临界阈值f_c越接近1。

为了清楚地展示f_c有多接近理论极限$f_c = 1$，我们来计算一下互联

网的f_c值。路由器层面的互联网满足$\frac{\langle k^2 \rangle}{\langle k \rangle} = 37.91$（表3-1）。将这个比例代入公式7.7，得到$f_c = 0.972$。因此，我们需要移除97%的路由器才能把互联网裂解成不相连的连通分支。在互联网的$N = 192\ 244$个路由器中，97%（也就是186 861个）的路由器同时出现故障的概率可认为是零。这就是互联网的拓扑结构对随机故障非常鲁棒的原因。

一般而言，如果一个网络的分解阈值高于随机网络的阈值（公式7.8），我们就说这个网络有增强的鲁棒性：

$$f > f_C^{\mathrm{ER}} \qquad (7.11)$$

随之而来，关于增强的鲁棒性会有如下的结论：

- 对于大多数$\langle k^2 \rangle$不同于$\langle k \rangle(\langle k \rangle + 1)$的网络而言，公式7.11成立。根据图3-8，几乎对于所有的参照网络，$\langle k^2 \rangle$都比随机网络中相应的期望值要高。因此，公式7.7给出的鲁棒性涉及我们实际关心的大多数网络。正如表7-1所示，公式7.11对于大多数参照网络都成立。

表7-1　在随机故障和蓄意攻击下的网络分解阈值

网络	随机故障（真实网络）	随机故障（随机网络）	攻击（真实网络）
互联网	0.92	0.84	0.16
万维网	0.88	0.85	0.12
电网	0.61	0.63	0.20
手机通话	0.78	0.68	0.20
电子邮件	0.92	0.69	0.04
科学合作	0.92	0.88	0.27
演员网络	0.98	0.99	0.55
引文网络	0.96	0.95	0.76
大肠杆菌代谢	0.96	0.90	0.49
酵母蛋白质相互作用	0.88	0.66	0.66

该表显示了在随机节点故障（第二列）和攻击（第四列）下，10个参照网络上的f_c的估计值。关于如何确定f_c的值，请参见进阶阅读7.E。第三列考虑了随机网络，其N和L与原始网络一样，但节点是随机连接的，即公式7.8计算得到的随机化网络的f_c^{ER}

。对于大多数网络，随机故障的f_c超过了相应随机网络上的f_c^{ER}，表明这些真实网络增强的鲁棒性，正如公式**7.11** 所表述的。然而，有3个网络却不是这样的：电网，因为它的度分布是指数分布的（图**7-31a**）；还有演员网络和引文网络，它们具有很高的$\langle k\rangle$。因此，尽管它们的$\langle k^2\rangle$值也很高，但高$\langle k\rangle$值削弱了$\langle k^2\rangle$的作用（公式**7.7**）。

- 公式**7.7** 说明，网络的度分布并不需要遵循严格的幂律才能有增强的鲁棒性。只要一个网络的$\langle k^2\rangle$高于规模相近的随机网络的$\langle k^2\rangle$的期望值即可。
- 网络的无标度特性不仅改变了f_c的值，还改变了在f_c邻域附近的临界指数γ_p、β_p和ν。关于这些指数和度指数γ的关系，请参考进阶阅读**7.A**。
- 增强的鲁棒性不仅体现在移除节点的情况下，还体现在移除链接的情况下（图**7-10**）。

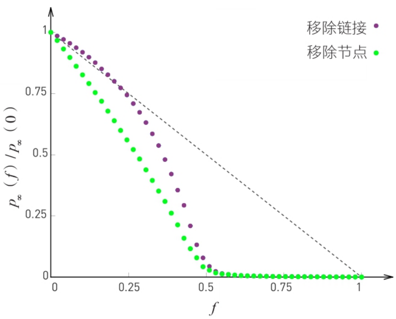


图**7-10**　鲁棒性与链接的移除

如果我们随机移除的不是节点，而是链接，情况会发生什么变化？对于随机链接移除和随机节点移除的情况，理论计算给出的临界阈值是相同的[7]，[8]。为了说明这一点，我们在$\langle k\rangle=2$的随机网络上比较随机节点移除和随机链接移除的影响。如图所示，网络在同一个临界值$f_c\approx0.5$上被裂解，不同的是这两条曲线的形状。事实上，去除占比为f的节点后剩下的巨连通分支，小于去除占比为f的链接后剩下的巨连通分支。这并不意外：因为平均每个节点有$\langle k\rangle$个链接，移除占比为f的节点等价于移除了占比为$f\langle k\rangle$的链接。这造成的损害显然高于只移除占比为f的链接。

总而言之，这一节我们考察了真实网络的一个基本特性：对随机故障的鲁棒性。公式**7.7** 告诉我们，网络的分解阈值取决于$\langle k\rangle$和$\langle k^2\rangle$这两个由网络的度分布唯一确定的参数。因此，在随机网络中，这一

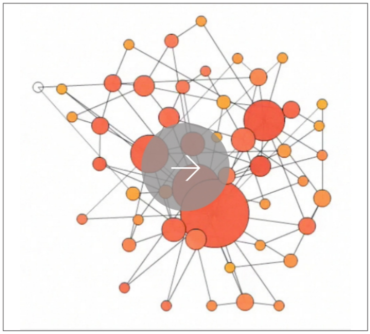
阈值是介于0和1之间的。然而，对于$\gamma < 3$的无标度网络而言，分解阈值降低到1。换言之，为了分解无标度网络，我们需要移除几乎所有的节点。而这正显示了这些网络对随机故障的极强的鲁棒性。

这种极端的鲁棒性是因为存在较大的$\langle k^2 \rangle$项。鉴于在大多数实际网络中，$\langle k^2 \rangle$都大于随机网络中的期望值，许多网络都具有增强的鲁棒性。存在这种鲁棒性是因为：随机故障主要影响的是众多的小度节点，而这些小度节点在维护网络完整性方面发挥的作用非常有限。

7.4 攻击耐受性

枢纽节点在维持无标度网络完整性中发挥着重要作用，这提出了下一个问题：如果我们移除节点的方式并不是随机的，而是针对枢纽节点的，会发生什么呢？也就是说，首先去掉网络中度最高的节点，然后去掉剩余节点中度最高的节点，这样重复下去，会发生什么呢？在正常情况下，节点按照这个特定的顺序被移除的可能性基本为零。实际上，这个移除顺序模仿了对网络的一种攻击。这种攻击假定我们了解网络的详细拓扑结构，有能力攻击枢纽节点，并且有动机故意破坏网络[1]。

移除单一枢纽节点是不可能分解一个网络的，因为网络中剩余的枢纽节点仍然可以将网络连接在一起。然而，去掉几个枢纽节点后，网络中的节点开始成片地脱落（在线资源**7.2**）。如果继续攻击，网络会迅速分解成很小的连通分支。



在线资源7.2　受到攻击的无标度网络

在攻击中，我们的目标是对网络造成最大的损害。我们可以通过首先移除度最大的节点，其次是次大的节点，以此类推。正如短片中所展示的，只移除几个枢纽节点就足够把无标度网络分解成不相连的连通分支。请把这个过程与在线资源7.1 所示网络在随机节点故障时难以被裂解的过程相比较。感谢王大顺提供可视化效果。

扫描二维码下载“湛庐阅读”App，回复书名，进一步了解受到攻击的无标度网络。

枢纽节点的移除对无标度网络的影响更加明显（图7-11　）：在随机的节点攻击下，网络中不存在分解的临界点。而在对枢纽节点的攻击中，临界点不仅存在，而且它的值非常低。因此，移除一小部分枢纽节点足以把无标度网络分解成众多小的连通分支。本节的目标是量化讨论网络在这种攻击下的脆弱程度。

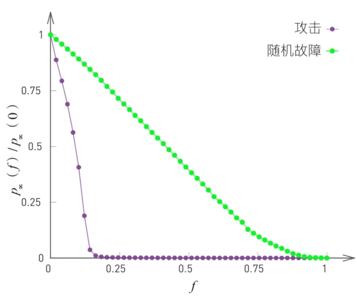


图7-11　受到攻击的无标度网络

此图描述的是，当一个无标度网络遭受攻击（紫色）和随机故障（绿色）时，一个节点属于巨连通分支的概率。攻击是按照节点度递减的顺序移除节点：从度最大的枢纽节点开始，然后是次大的，以此类推。故障是随机选择节点，顺序与节点的度无关。此图展示了无标度网络受到攻击时的极端脆弱性：f_c 很小，意味着只要移除少数几个枢纽节点就可以分解网络。初始网络度指数为$\gamma=2.5$，$k_{min}=2$，$N=10\ 000$。

攻击下的临界阈值

对无标度网络的攻击有以下两个结论（图7-11）：

- 临界阈值f_c 小于1意味着，在攻击无标度网络时，只需要移除一部分枢纽节点即可使其分解。
- 所观察到的f_c 非常低，这表明我们只需要移除很小一部分枢纽节点，就可以使网络失效。

为了量化这个过程中，我们需要解析计算网络受到攻击时的f_c值。我们考虑移除枢纽节点给网络带来的两点改变[9]：

- 网络最大的度从k_{max} 变成了k'_{max}，因为度大于k'_{max} 的所有枢纽节点都被移除了。
- 度分布从p_k 变成了$p'_{k'}$，因为枢纽节点被移除后，其上的链接也随之移除，原本与之相连的节点的度发生变化。

融合以上两个因素，我们可以把攻击问题对应到上一节讨论的鲁棒性问题上。换言之，我们可以把对网络的攻击看作在最大度为k'_{max}度分布为$p'_{k'}$ 的另一个网络上的随机节点移除。通过计算可知，无标度网络受到攻击时的临界阈值f_c 是以下方程的解[9]，[10]（进阶阅读7.F）：

$$f_c^{\frac{2-\gamma}{1-\gamma}}=2+\frac{2-\gamma}{3-\gamma}k_{min}\left(f_c^{\frac{3-\gamma}{1-\gamma}}-1\right) \tag{7.12}$$

图7-12 展示了公式7.12 的数值解关于度指数γ 的函数关系，可以得出以下结论：

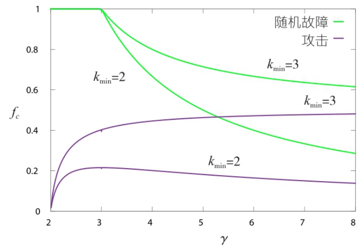


图7-12　攻击下的临界阈值

在$k_{\min}$ =2或3的无标度网络中，分解阈值f_c 对度指数γ 的依赖关系。攻击的紫色曲线由公式7.12 给出，随机故障的绿色曲线由公式7.7 给出。

- 在随机故障中，临界值f_c 随γ 的增加而单调减小。但在攻击下，f_c 并非单调的：当γ 较小时，f_c 随γ 的增加而增加，但当γ 较大时，f_c 随γ 的增加而减小。
- 攻击下的f_c 总是小于随机故障的f_c 。
- γ 值较大的无标度网络的表现与随机网络相似。因为随机网络没有枢纽节点，所以攻击的影响与随机节点移除的影响并没有什么不同。因此，当γ 较大时，随机移除的阈值与攻击的阈值会降低到同一个值。事实上，如果$\gamma \to \infty$ ，我们有$p_k \to \delta(k - k_{\min})$，意味着所有节点具有相同的度$k_{\min}$ 。于是，在$\gamma \to \infty$的极限情况下，随机故障和攻击是不可区分的，可得出以下公式：

$$f_C \to 1 - \frac{1}{(k_{\min} - 1)} \quad (7.13)$$

- 如图7-13 所示，一个随机网络在随机故障和攻击中都存在介于0和1之间的渗流阈值，这与图7-12 及公式7.13 在γ 较大的情况下一致。

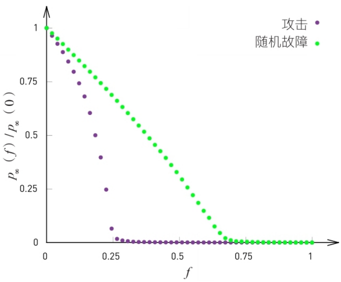


图7-13　随机网络中的攻击与故障

在随机网络中，占比为f 的节点被移除后，巨连通分支所占的比例。移除可以是随机移除（绿色），或者按照节点的度降序移除（紫色）。这两条曲线都表明存在一个介于0和1之间的阈值f_c ，不同于无标度网络在随机故障下$f_c \rightarrow 1$的情况。该结果来自$N = 10\ 000$且$\langle k \rangle = 3$的随机网络上的模拟。

机场的类比有助于我们了解无标度网络面对攻击时的脆弱性：只要关闭两个大型机场，例如芝加哥的奥黑尔机场或亚特兰大国际机场，几个小时内就会产生头条新闻，全美所有的旅行计划都需要随之变更。如果有一系列事件导致亚特兰大、芝加哥、丹佛和纽约机场这些最大的枢纽节点同时关闭，那么北美大陆的航空旅行将在几个小时内停止。

总而言之，虽然随机节点故障不能分解无标度网络，但攻击枢纽节点可以轻易将其摧毁。这一脆弱性对互联网来说是个坏消息，这表明，在本质上，互联网面对蓄意攻击时是脆弱的。不过，这在医学上是个好消息，因为这表明，细菌面对去除枢纽蛋白时是脆弱的，这给我们提供了渠道，可以设计药物来杀死不需要的细菌。

边栏7.2

保罗·巴兰与互联网

1959年，加利福尼亚州的兰德智库公司（RAND）派当时年轻的工程师保罗·巴兰开发一个通信系统，要求该系统在苏联核攻击后仍能坚持工作。因为核打击会使爆炸范围内的所有设备无法工作，巴兰需要设计一个系统，使爆炸范围外的用户仍能彼此联系。巴兰

把那个时代的通信网络描述为“由一些星型网络连接成一个更大的星型网络的层次结构”，这正是对今天所谓的无标度网络的早期描述 [11] 。巴兰的结论是，这种拓扑结构过于中心化，在攻击下不可行。巴兰也舍弃了图**7-14a** 所示的中心辐射拓扑。他指出：“中心化的网络显然是脆弱的，因为破坏一个中央节点就可以破坏终端站之间的通信。”

巴兰认为，能在攻击下存活的理想的网络结构是一个分布式的网格状网络（图**7-14c** ）。这个网络足够冗余。因此，即使它的一些节点发生故障，还有其他路径将剩下的节点连接起来。巴兰的想法未被军方采纳。因此，10年后诞生的互联网依赖于分布式网络协议，允许每个节点自行决定链接到哪里。这种去中心化的思想决定了互联网会成为无标度网络，而不是像巴兰所设想的均匀分布的网格状拓扑。

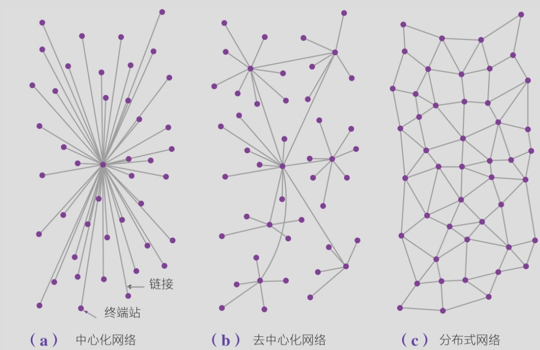


图7-14 巴兰设想的网络

保罗·巴兰在1959年设想的几种可能的通信网络。 [11]

7.5 级联故障

在这一章之前的讨论中，我们一直假设每个节点故障是一个随机事件，因此网络中各个节点发生故障的概率是相互独立的。但在现实中，网络中的每个节点的活动都与相邻节点的活动有关。因此，一个

节点的故障可以导致相邻节点的故障（图7-15　）。我们先来看几个例子。



图7-15　多米诺效应

多米诺效应是指一块多米诺骨牌倒下引起一连串多米诺骨牌接连倒下的现象。这个术语经常被用来指这样一类事件：一个局部变化引发了一系列事件，进而传播到整个系统。因此，多米诺效应也许是能说明本节主题级联故障的最简单的例子。

- 停电（电网）

一个节点或一个链接发生故障之后，电流会瞬间在电网其余部分重新分配。例如，1996年8月10日，俄勒冈处于高温天气，一条承载1　300兆瓦的电线，碰到树枝后断裂。因为电无法存储，其承载的电流自动转移到电压稍低的另外两条输电线路上。由于后者并非设计用于承载额外电流，它们也随之发生故障。几秒钟后，多余的电流导致13个发电机发生故障，最终导致波及美国11个州和加拿大两个省的大规模停电[12]。

- 拒绝服务攻击（互联网）

如果一个路由器无法传输它接收到的数据包，互联网协议会通知相邻的路由器把数据包重新分配到其他路径，以绕过这个有故障的路由器。因此，一个路由器的故障会增加其他路由器上的流量，有可能会在整个互联网上引发一系列拒绝服务攻击[13]。

- 金融危机

级联故障在经济系统中很常见。例如，2008年美国房价下跌的影响通过金融网络的链接蔓延，导致了级联式的银行、公司，甚至国家

陷入危机[14-16]，最终引起了自20世纪30年代大萧条以来最严重的全球金融风暴。

虽然这些例子来自不同的领域，但它们有几个共同的特点。首先，最初的故障对网络结构的影响非常有限。其次，最初的故障并没有停留在局部，而是沿着网络的链接传播，引发了更多的故障。最后，许多节点无法履行正常功能。因此，这些系统中都发生了级联故障，这是大多数网络都存在的危险现象[17]。在这一节中，我们将从经验角度讨论级联故障中的模式。在下一节，我们将对级联故障建模。

经验性结果

电网、信息系统、地壳构造运动中的级联故障都有详细的记录，可以查到级联故障的频率和强度的具体统计数据。

- 停电

电站故障、传输电线损坏、短路等原因都可能导致停电。当电网中某个组件超过了运行限制时，它就会自动从网络中断开以自我保护。这会使原本由该故障组件承载的电力转移到其他组件上，改变电流的潮流（power flow）、频率、电压和相位，以及控制、监控、预警系统的运行。这些变化也可能反过来让其他组件从网络中断开，引发雪崩式的故障（图**7-16**）。



图7-16　2003年北美东北地区的停电范围

2003年8月14日下午4:10发生的停电事件是北美史上最大规模的停电事件之一。它的原因来自第一能源公司在俄亥俄州一间控制室的里一个预警系统的软件错误。由于没有收到警报，电力运营商不知道在一根过载高压线碰到树后，需要在网络中重新分配功率。因此，一个本可以在局部被控制的故障引发了级联故障，致使265家发电厂的超过508个发电单位被关闭，导致安大略省大约1 000万人、美国8个州超过4 500万人失去供电。此图展示了在2003年8月14日的停电事件中受影响的州。关于这次停电的卫星图片，请见图**0-1** 。

停电规模通常以未能供应的电量来度量。图**7-17a** 显示了1984年和1998年之间，北美历次大规模停电中未能供应的电量的概率分布$p(s)$。电气工程师将该分布以幂律近似为以下公式[18] ：

$$p(s) \sim s^{-\alpha} \tag{7.14}$$

其中雪崩指数α 在不同国家的取值可见表**7-2** 。这种分布的幂律特性表明，尽管大多数的停电范围是相当小的，只影响少数消费者，然而也存在着偶尔发生的大规模停电，使数百万消费者失去供电。

- 信息的级联传播

从电子邮件到社交网站，现代通信系统使信息可以沿着社交网络的链接进行级联式的传播。信息传播的过程通常会留下电子记录，这些平台让研究人员能够检测到其背后的级联现象。

很多研究关注提供微型博客服务的Twitter。在Twitter上，我们可以通过抓取关注者的网络来重建“谁关注谁”的网络。由于用户经常使用短网址来分享网页内容，我们可以记录每一次扩散／分享的过程。一项研究追踪了两个月内的7　400万个网址扩散事件。每一次扩散都是从某一个种子节点发布网址开始，经过级联式的转发，直到级联传播结束（图7-18 ）。图7-17b 表明，所观察到的级联传播的规模符合幂律（公式7.14 ），其雪崩指数为$\alpha \approx 1.75$ [19] 。幂律表明，绝大多数发布的网址完全没有得到扩散，而只有一小部分网址可以被转发上千次，前者的结论也可以从平均转发量的规模只有$(s)=1.14$看出来。

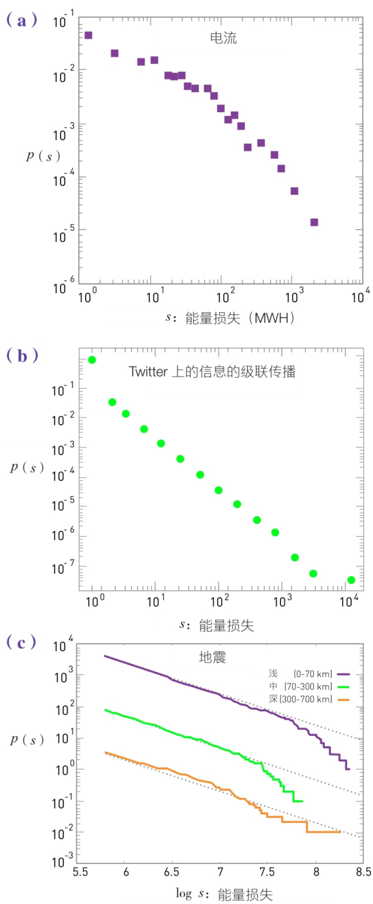


图7-17　级联规模的分布

（a）北美电力可靠性委员会记录的1984年至1998年之间北美所有由停电导致的能源损失的分布。这个分布与公式7.14 十分契合。不同国家的指数列在了表7-2中。[18]

（b）Twitter上的信息级联传播的规模分布。虽然Twitter大部分信息都无人关注，但有很少一部分作息被转发了数千次。总体来说，转发量可以被公式7.14　很好地近似，幂律指数$\alpha \backsimeq 1.75$。[19]

（c）1977年至2000年之间记录的地震振幅分布。虚线是地质学家描述该分布时给出的幂律拟合公式7.14 。地震的振幅在横坐标上用s 的对数表示，s 是检测到的地震波的振幅。[20]

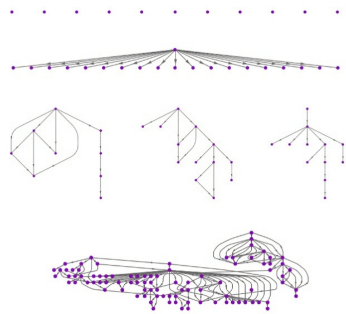


图7-18　信息的级联传播

Twitter上的信息级联传播的例子。节点代表Twitter用户。顶部的节点对应第一个发布某个短网址的用户。链接表示转发。这些级联现象捕捉到了信息雪崩的异质性：大多数Twitter信息完全没有任何转发，在图中只是单个节点。然而，如底部子图所示，有一些内容却呈现出雪崩式的大规模转发。[19]

- 地震

地质断层面是不规则的且有黏性的，这使得它们之间很难相互滑动。一旦断层闭锁，构造板块的持续相对运动就会积累越来越多的应变能。当压力积累到足够克服断层之间的粗糙阻力时，断层面就会忽然滑动，并且释放所储存的能量，引发地震。地质断层的自然破裂、火山活动、山体滑坡、矿场爆炸甚至核试验，都可能引起地震。

每年大约有50万次的地震被仪器检测到。其中，只有约10万次强烈到能被人体感知。地震学家用幂律来近似拟合地震振幅的概率分布

（公式**7.14**），得到$\alpha\approx 1.67$（图**7-17c**）[20]。

人们很少把地震视为一种明显的网络现象，这是因为很难找到由相互依赖关系构成的引发地震的具体网络。然而，地震中的级联后果与网络中的级联现象有许多相似性，暗示它们背后存在相同的机制。

停电、信息级联传播和地震背后的幂律分布（公式**7.14**）表明，大多数的级联故障都是比较小的。例如这些小的级联故障只涉及几所房子的停电、大多数用户不感兴趣的Twitter信息、微弱到需要用敏感仪器才能检测到的地震。公式**7.14** 指出与这些众多的小事件并存的是少数的大事件。这些大的级联事件包括：2003年北美大停电（图**7-16**）、转发1 399次的Twitter信息——《伊朗大选危机：10个难以置信的YouTube视频》[21]、2010年1月超过20万人遇难的海地地震。有趣的是，由电气工程师、媒体研究人员和地震学家报告的雪崩指数惊人地接近，都介于1.6和2之间（表**7-2**）。

表**7-2** 真实网络中的雪崩指数

来源	雪崩指数	级联的度量对象
电网（北美）	2.0	电力
电网（瑞典）	1.6	能源
电网（挪威）	1.7	电力
电网（新西兰）	1.6	能源
电网（中国）	1.8	能源
信息传播	1.75	转发量
地震	1.67	地震波

各种级联的幂律分布（公式**7.14**）的雪崩指数，包括各个国家在停电中的电量损失[18]、Twitter上的信息传播[19]、地震的规模[20]。第三列表示级联规模s的度量对象，对应停电中未能供应的电力和能源、某一条Twitter信息的转发量以及地震波的振幅。

其他许多场景中也存在级联故障：

- 恶劣天气或机械故障会沿着航线时刻表级联式地蔓延，导致多个航班延误和数千名旅客滞留（边栏**7.3**）[22]。

边栏7.3

航班中的级联式拥堵现象

美国每年因航班延误造成的经济损失超过400亿美元 [28] 。这些损失包括延误造成的额外运营、乘客时间的浪费、生产力的下降、错失的商业和娱乐机会。航班延误是指计划起飞时间与实际起飞或到达时间之间的差异。航空时间表规划了连续航班之间的缓冲期以应对短期延误。当延误超过缓冲期，使用同一架飞机、同一批机组人员或同一个登机口的后续航班都会被推迟。因此，延误以类似级联的方式在航空网络中传播。

虽然2010年的航班大多数是准时的，依然有37.5%的航班的起飞或到达时间晚于预计 [22] 。延迟的时间分布服从公式**7.14** ，这表明，尽管大多数航班只推迟了几分钟，也存在少数航班被推迟了几个小时的情况。这些长时间的延误表现出了相互关联的模式，标志着在航空运输系统中发生了级联式的拥堵现象（图**7-19** ）。

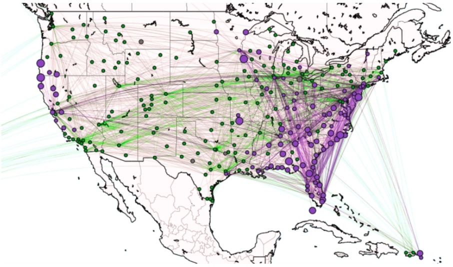


图**7-19**　拥堵机场的簇

美国航空地图上的拥堵机场（紫色节点）和正常运输的机场（绿色节点）。连线对应2010年3月12日各个机场之间的直飞航班。拥堵机场构成的簇表明，航班延误并非彼此独立，而是呈现出航空网络上的级联现象。 [22]

- 一个物种的消失会通过食物网在生态系统中产生级联效应，使许多物种灭绝，并改变其他物种的栖息环境 [23-26] 。
- 供应链中某个组件短缺会导致整条供应链无法正常工作。例如，2011年在泰国发生的洪水导致了汽车零部件

长期短缺，扰乱了全球上千家汽车工厂的生产链。因此，洪水的危害不仅限于被淹没的工厂，还导致了世界范围内高达200亿美元的保险索赔。

总而言之，在许多截然不同的系统中，我们都观察到了级联效应。它们的规模分布可以由幂律很好地近似。这意味着，大多数的级联事件非常小，以至于我们注意不到；然而，也存在着少数具有全球性影响力的重大事件。在下一节中，我们将理解这些现象的起因，并建立模型来再现其特征。

7.6 对级联故障建模

级联现象的发生与很多因素有关，包括级联传播所在的网络结构、级联传播的机制、每个个体产生故障的条件。经验表明，虽然这些因素各不相同，但我们观察到的级联规模的分布有一些不拘于具体系统的共性。在这一节中，我们将理解这些级联现象背后的机制，并解释雪崩规模分布的本质。

为了了解级联现象的动力学本质，研究人员提出了很多模型[18]，[29-35]。虽然这些模型描述具体现象的真实程度不尽相同，但它们都指出发生级联现象的系统具备以下三个关键因素：

（1） 系统可以被描述为一个存在流量的网络，就像电网中的电流或者通信网络中的信息流。

（2） 每个连通分支都存在一种局域故障规则，它决定了一个连通分支何时会参与到级联现象中去。参与的形式可能是发生故障（比如电网、地震）或者选择传播某一条信息（比如Twitter）。

（3） 在连通分支遇到故障或者被激活时，系统有重新分配流量的机制。

下面，我们将讨论两个级联故障模型，它们可以预测不同抽象程

度的级联故障的特征。

故障传播模型

尽管故障传播模型最初被用于建模想法和观念的传播[30]，它也经常被用来描述级联故障[35]。该模型定义如下：

我们先来看一个任意度分布的网络，每个节点都有一个代理（agent）。代理i可以存在两种状态：状态0（活跃或者正常）和状态1（不活跃或者故障）。每个代理i都有一个故障阈值$\varphi_i = \varphi$。

在开始的时候，所有的代理都处于正常状态0。在$t = 0$时，一个代理的状态变为1，代表最初的一个组件故障，或者最初发布的一条新信息。在之后的每一步中，我们随机地选取一个代理，并根据阈值法则来更新它的状态：

- 如果该代理i的状态为0，那么它将根据其邻居的状态做出更新：如果k_i个邻居中至少有ϕ比例的邻居的状态为1，那么代理i将改变状态为1，也就是说，代理i也发生故障；否则，代理i保持状态为0。
- 如果该代理i的状态为1，那么它的状态不变。

换言之，当一个正常的节点i有占比为ϕ的邻居发生故障时，它就会改变状态。根据局域网络拓扑不同的特征，最初的扰动可能会立刻消失，而不引发其他节点的故障，也可能如图**7-20a**、**b**所示引发多个节点的故障。模型实验显示了三种情形，它们有截然不同的雪崩特征（图**7-20c**）。

- 次临界情形

如果$\langle k \rangle$较高，那么改变单个节点的状态不太可能让其他节点达到临界值，因为正常节点有很多正常的邻居。在这种情形下，级联会很快消失，级联规模服从指数分布。因此，在这样的系统内不会发生全局性的大规模级联故障（图**7-20c**、**d**中的蓝色符号）。

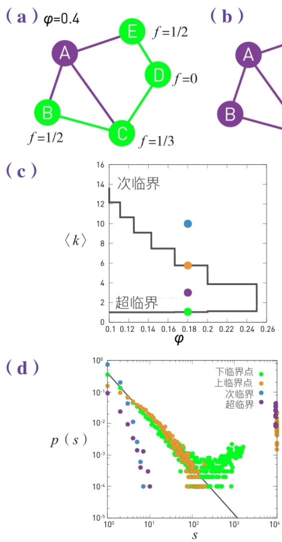


图7-20　故障传播模型

（a）（b）这两个图描述的是一个小网络的级联故障的发展过程。网络中每个节点的故障阈值都为ϕ =0.4。最初所有节点都处于状态0，用绿色圆圈表示。首先将节点A的状态更改为1（紫色），那么它的邻居B和E将有占比为f =1/2＞0.4的邻居，其状态为1。因此，它们也会发生故障，把状态改变为1，如图b所示。在下一步中，C和D也将发生故障，因为它们都有f ＞0.4的邻居发生了故障。因此，级联会贯穿整个网络，最终的规模为s =5。读者可以自行验证，如果首先改变状态的节点是B，它将不会引发雪崩。

（c）此图是故障传播模型的相图，表示为阈值函数ϕ 和所在网络平均度$\langle k \rangle$构成的空间。在($\langle k \rangle$,ϕ)空间中用实线包围的区域，级联可以在随机图上传播。

（d）此图描述的是在N =10 000、ϕ =0.18的网络上，当$\langle k \rangle$ =1.05（绿色）、$\langle k \rangle$ =3（紫色）、$\langle k \rangle$ =5.76（橙色）和$\langle k \rangle$ =10（蓝色）时各自的级联规模分布。在下临界点，我们观察到指数为α =3/2的幂律分布p (s)。在超临界情形下，除了几个小规模的级联，大部分级联都是全局性的。在上临界点处和次临界情形下，我们只能观察到小规模级联。[30]

- ## 超临界情形

如果$\langle k \rangle$ 较低，那么改变单个节点的状态可能会让几个邻居节点达到临界值，进而引发全局性的级联故障。这种情形下，扰动可以引发大范围的系统故障（图**7-20c**、**d** 中的紫色符号）

- 临界情形

在次临界与超临界的交界，级联规模可大可小，分布范围很广。数值模拟表明，如果网络是随机的，那么此时的级联规模服从指数$\alpha=3/2$的幂律分布（如公式**7.14** 所示）（图**7-21d** 中的绿色和紫色符号）。

分支模型

故障传播模型的复杂性使其很难用于分析级联的标度。为了理解$p(s)$的幂律本质并计算雪崩指数α，我们转而考虑分支模型。分支模型是能够抓住级联事件基本特征的最简单的模型。

分支模型基于这样的观察：每一次级联故障都是一次分支过程。我们将最先发生故障并引发级联故障的节点称为树的根节点。树的分支就是由这个初始故障引发后续故障的节点。例如图**7-20a**、**b** 所示，节点A的故障引发了级联故障，因此A是树的根节点。节点A的故障导致了节点B和节点E的故障，这两个节点就是树的两个分支。接下来，节点E的故障导致了节点D的故障，节点B的故障导致了节点C的故障（图**7-21a** ）。

分支模型捕捉到了级联传播的基本特征（图**7-21** ）。该模型从单个活跃节点开始，在之后的每一个步中，每个活跃节点产生k个后代，其中k由概率分布p_k生成。如果一个节点生成了$k=0$，那么这个分支就结束了（图**7-21b** ）。如果一个节点生成了$k>0$，它将会产生k个新的活跃后代。当所有活跃节点的分支都结束时，这个树的规模就是级联事件的规模（图**7-21c** ）。

分支模型给出的相与级联故障模型观察到的三种情形相同。相只由$\langle k\rangle$决定，也就是只由p_k决定：

- 次临界情形：$\langle k\rangle<1$

当$\langle k\rangle<1$时，平均每个分支只有不到一个后代。因此，每一棵树都会很快结束（图**7-21d** ）。在这种情形下，级联规模服从指数分

布。

- 超临界情形：$\langle k \rangle > 1$

当$\langle k \rangle > 1$时，平均每个分支会有多于一个后代。因此，树会无穷增长（图**7-21e**）。因此，此时的级联全都是全局性的。

- 临界情形：$\langle k \rangle = 1$

当$\langle k \rangle > 1$时，平均每个分支都有一个后代，因此，有些树会长到很大，有些会很快结束（图**7-21e**）。数值模拟表明，此时级联规模服从幂律分布（公式**7.14**）。

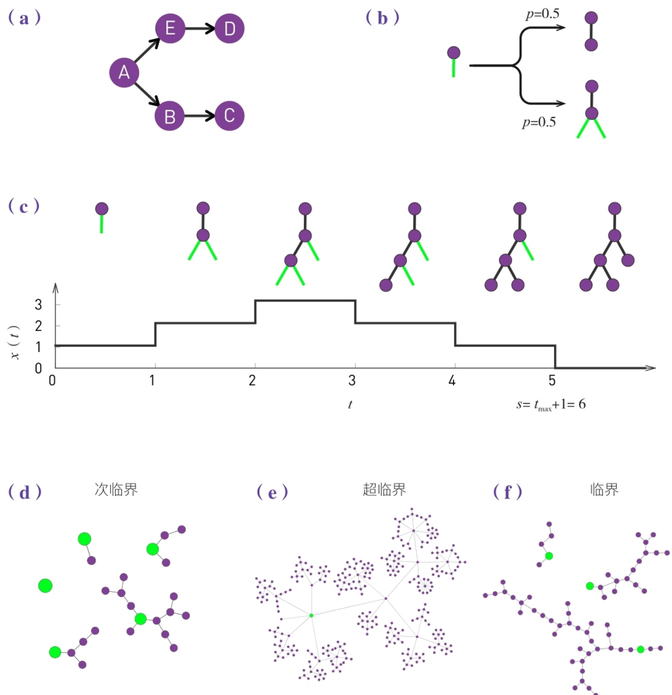


图**7-21**　分支模型

（**a**） 此图用分支过程来表述图**7-20a**、**b** 所示的故障传播。扰动从节点A的故障开始，这一故障改变了B和E的状态，它们又分别改变了C和D的状态。

（**b**） 此图描述了一个简单的分支过程。每一条活跃链接（绿色）会以$p_0=1/2$的概率（上图）变为不活跃，或者以$p_2=1/2$的概率（下图）产生2个新的活跃链接。

（**c**） 为了解析计算$p(s)$，我们把分支过程映射到扩散问题。为此，我们给出活跃节点的数目$x(t)$关于时间t的函数。非零的$x(t)$意味着级联还在继续。当$x(t)$变为零时，不再有活跃节点，级联结束。如图所示的例子中，级联结束发生在$t=5$时，因此，级联的规模为$t_{max}+1=6$。

分支过程可以映射到一维随机游模型中，这可以帮助我们计算级联的雪崩指数。一个分支过程始于一条半链接（只有一端活跃的链接）。当一个活跃节点转为不活跃时，活跃节点数减少，即$x\to x-1$。当活跃节点产生分支时，会产生两个新的活跃节点，即$x\to x+1$。这一过程把级联的规模s映射到了随机游走从$x=1$开始到首次返回$x=0$所用的时间。随机游走理论对这个问题已有不少研究，指出返回时间服从指数为3/2的幂律分布[32]。若分支过程对应度分布为p_k的无标度网络，级联指数则依赖于γ，如图**7-22**所示。

（**d**）～（**f**） 这三个图描述的是，由分支模型在次临界（图d）、超临界（图e）和临界（图f）情形下产生的典型级联现象。在每一个级联中，绿色节点代表树的根节点，也就是第一个扰动。图d和图f中展示了多棵树，而图e中只展示了一棵树，因为后者的每一棵树（级联）都会无穷增长。

分支模型可以用于求解，因而对于任意p_k，我们可以计算出级联规模的分布。如果p_k的分布是指数有界的，例如分布的尾部是指数分布，则可以计算得到$\alpha=3/2$。然而，如果p_k是无标度的，那么级联的雪崩指数取决于p_k的幂律指数γ，如下图**7-22**所示[32]，[33]：

$$\alpha=\begin{cases}\dfrac{3}{2}, \gamma\geqslant 3\\ \dfrac{\gamma}{\gamma-1}, 2<\gamma<3\end{cases} \tag{7.15}$$

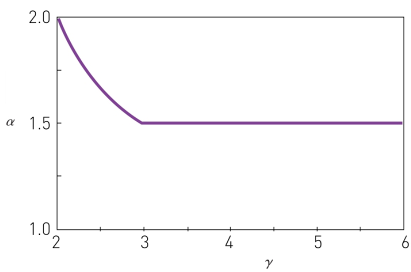


图**7-22** 雪崩指数

此图描述的是，公式**7.15** 中指出的雪崩指数α 对所在网络的度指数γ 的依赖关系。曲线表明，当$2<\gamma<3$时，雪崩指数和度指数有关。然而，当$\gamma\geq 3$时，雪崩与随机网络一致，为$\alpha=3/2$。

我们带着这一结果重新审视表**7-2**，发现经验观察到的雪崩指数都在1.5和2之间，与公式**7.15** 吻合。

总而言之，本节讨论了两个可以描述级联故障动力学特征的模型：故障传播模型与分支模型。在文献中，我们还可能会遇到用于描述电网故障的过载模型 [18] 和用于描述临界状态下级联故障的沙堆模型 [31]，[32]。还有一些其他的模型可以考虑节点和链接不同的承载能力 [34]。这些模型的现实意义各不相同，它们可以允许不同性质和数量的参数调节。但是，它们都指出存在一个临界状态，在这个状态下雪崩规模服从幂律分布。雪崩指数α 由级联传播所在网络的度指数唯一决定。有着截然不同传播动力学性质和故障机理的模型得到了相同的标度性质和雪崩指数，说明它们背后的现象是普适的，即与模型无关的。

7.7 构筑鲁棒性

能否提高一个网络的鲁棒性？在这一节中，我们将讨论影响鲁棒性的因素，从而设计能同时抵御随机故障和攻击的网络。我们也会讨论如何让级联故障停止，使系统的动态鲁棒性提高。最后，我们将这些工具应用到电网上，把鲁棒性与可靠性关联起来。

设计鲁棒的网络

设计可以同时抵御攻击和随机故障的网络似乎是一个自相矛盾的愿望 [36-39]。例如，图**7-23a** 所示的中心辐射网络对随机故障具有鲁棒性，因为只有当它的中心节点发生故障时，才能把网络分解成孤立的连通分支。因此，随机故障将网络碎片化的概率是$1/N$，当N 很大时，这个概率可以忽略不计。与此同时，这个网络面对攻击时很脆

弱，因为移除单个节点，即中心枢纽节点，就可以将网络分解成孤立的节点。

我们可以把这个网络的外围节点连接起来，以此来增强网络对攻击的耐受性（图**7-23b**），这样网络不再会因为移除中心节点而成为碎片。然而，增强鲁棒性的代价是链接数量需要加倍。假设建立和维护一个网络的成本与它的平均度$\langle k\rangle$成正比，那么图**7-23b**中所示的网络的成本是24/7，是图**7-23a**中所示网络的成本的2倍。网络成本的增加促使我们更仔细地提出这个问题：是否可以最大限度地提高网络对随机故障的鲁棒性和对攻击的鲁棒性，而不增加成本？

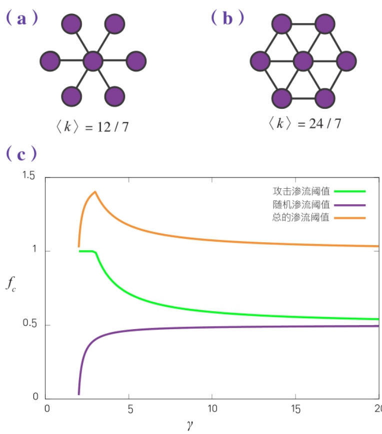


图**7-23**　增强鲁棒性

（**a**）　中心辐射网络对随机故障是鲁棒的，但对移除中心枢纽节点的攻击的耐受性很弱。

（**b**）　把小度节点连接起来后，强化后的网络对针对性攻击有了更高的耐受性。同时，以$\langle k\rangle$度量的成本也升高了。

（**c**）　此图描述的是，$k_{\min}=3$的无标度网络上随机渗流阈值f_c^{rand}、针对性攻击的渗流阈值f_c^{rand}和总的渗流阈值f_c^{tot}关于度指数γ的函数。

网络对随机故障的鲁棒性可以由它的渗流阈值f_c来描述，这个阈值表示为了分解网络，我们必须移除的节点的比例。为了增强网络的鲁棒性，我们需要提高f_c的值。根据公式**7.7**，f_c只取决于$\langle k\rangle$和$\langle k^2\rangle$。因此，在保持成本$\langle k\rangle$不变的情况下，最大化f_c要求最大化$\langle k^2\rangle$。这

可以通过双峰分布来实现：网络只有两类节点，一类节点的度都是$k_{\min}$，另一类节点的度都是$k_{\max}$（图**7-23a**、**b**）。

如果想优化网络拓扑结构，使其可以同时抵御随机故障和攻击，我们需要一种拓扑结构来最大化下列公式（图**7-24c**）：

$$f_c^{\text{tot}} = f_c^{\text{rand}} + f_c^{\text{targ}} \quad (7.16)$$

结合解析讨论和数值模拟，可以证明，双峰分布也使这个值达到最大[36-39]：

$$p_k = (1-r)\delta(k-k_{\min}) + r\delta(k-k_{\max}) \quad (7.17)$$

这个分布描述了这样一个网络，占比r的节点的度都为$k_{\min}$，另外，占比$(1-r)$的节点的度都为$k_{\max}$。

正如进阶阅读**7.G**中所示，当$r=1/N$时，f_c^{tot}取得最大值。也就是说，只有一个节点的度为$k_{\max}$，其他所有节点的度都为$k_{\min}$。在这种情况下，$k_{\max}$与网络规模的关系为：

$$k_{\max} = AN^{2/3} \quad (7.18)$$

换言之，对随机故障和攻击都鲁棒的网络应该只有一个度为$k_{\max}$的枢纽节点，而其他节点的度都为$k_{\min}$。很明显，这个中心辐射的拓扑结构对随机故障是鲁棒的，因为移除中心枢纽节点的概率是$1/N$，当N很大时，这个概率很小。

虽然这个网络看上去对移除枢纽节点的攻击表现得非常脆弱，但实际上并非如此。事实上，枢纽节点和其他很多度为$k_{\min}$的节点共同承担了维系网络最大连通分支的职责。当$k_{\min}>1$时，这些节点自己就组成了巨连通分支。因此，虽然移除$k_{\max}$的枢纽节点会造成一次性的重大损失，但其他的低度节点对后续攻击仍是鲁棒的（图**7-24c**）。

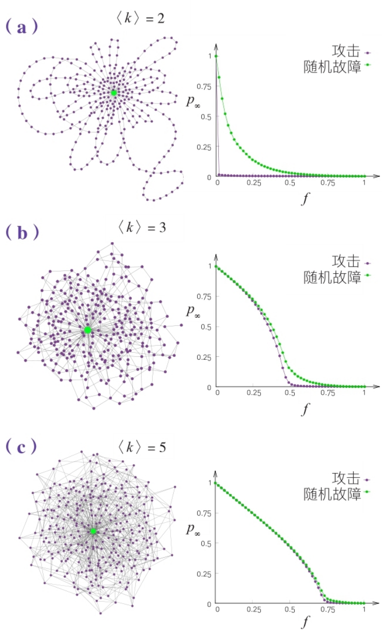


图7-24　优化对攻击和随机故障的耐受性

图中展示了由公式**7.16**和公式**7.17**给出的最优网络结构。这个网络有一个枢纽节点，它的度由公式**7.18**决定；其余节点有相同的度$k_{\min}$，由$\langle k \rangle$决定。左侧子图展示了$N=300$时网络的拓扑结构，右侧子图是$N=10\ 000$时的故障/攻击曲线。

（**a**）当$\langle k \rangle$较小时，枢纽节点负责维系网络。一旦我们移除中央的枢纽节点，网络就会分解成碎片。因此，攻击曲线和故障曲线相去甚远，表明网络对随机故障具有很强的鲁棒性，对攻击很脆弱。

（**b**）当$\langle k \rangle$较大时，即使在网络中缺少中央枢纽节点，也会出现巨连通分支。因此，尽管枢纽节点提升了网络对于随机故障的鲁棒性，但枢纽节点并非必不可少。在这种情况下，攻击阈值f_c^{targ}和故障阈值f_c^{rand}都较大。

（**c**）当$\langle k \rangle$非常大时，故障和攻击的曲线已经不可区分。这意味着网络对于攻击和随机故障的反应没什么不同。在这种情况下，网络即使没有枢纽节点，也仍然能完好地连接。

边栏**7.4**

阻止级联故障

我们可以避免级联故障吗？我们的第一反应可能是添加新的链

接以强化网络。但强化网络的问题是，在大多数实际系统中，建立一个新链接所需的时间远远长于级联故障的时间尺度。例如，限于监管、资金和法律的阻碍，在电网中建设新的输电线可能需要20年。相比之下，一次级联故障可以在几秒钟内横扫整个电网。

与直觉相反的是，选择性删除节点和链接可以减弱级联故障的后果[42]。为此，我们注意到每一个级联故障有两部分：

（1）　初始故障，也就是第一个节点或链接的故障，是后续级联故障的源头。

（2）　传播，初始故障导致其他节点发生故障，并开始通过网络引发级联效应。

通常情况下，（1）和（2）之间的时间间隔比强化网络所需的时间短得多。然而，模拟表明，如果我们在初始故障（1）之后、故障传播之前，故意移除一些其他节点，便可以缩小级联的规模。虽然刻意删除节点或链接是对网络的进一步破坏，但精心选择所移除的部分可以抑制级联传播[42]。模拟结果表明，为了限制级联的规模，我们必须在初始故障附近删除负载小的节点和负载过大的链接。该机制与消防员用的方法相似，在防火道上点一把可控的火，以耗尽山火途中的可燃物。

拉撒路效应（Lazarus effect）是运用这一方法带来的惊人效果。拉撒路效应是指，若想让一个无法生长和繁殖的已“死”细菌复活，只需删除几个精心挑选的基因（图7-25）[43]。因此，用一个违反直觉的方式，受控的损害反而对网络有利。

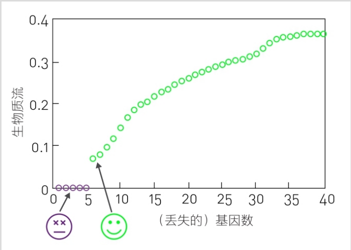


图7-25　拉撒路效应

细菌的生长率由它生产生物质（biomass）的能力决定。生物质是细菌用来制造细胞壁、DNA及其他成分的分子。如果一些关键基因缺失，细菌就无法继续产生所需要的生物质。如果细菌不能繁殖，就意味着死亡。如果缺失某个基因会导致生物质流为零，这样的基因则被称为必需基因。

图中曲线显示了大肠杆菌的生物质流。大肠杆菌是生物学家在研究中经常用到的一种细菌。图中大肠杆菌发生了基因突变，一个必需基因丢失，因此纵轴所示的生物质流为零，此时大肠杆菌无法再繁殖。然而，正如图中所示，如果再移除另外5个基因，我就可以重新激活生物质流。就这样，我们反直觉地移除更多的基因，复活了一个已死的有机体。这个现象被称为拉撒路效应[41]。

案例研究：估计鲁棒性

欧洲电网集合了20多个国家的电网，包括超过3 000个发电机和变电所（节点），以及20万千米长的输电线（图**7-26a～d**）。网络度分布可以近似为下列公式（图**7-26e**）[40]，[41]：

$$p_k = \frac{e^{-k/\langle k\rangle}}{\langle k\rangle} \quad (7.19)$$

这表明它的拓扑结构由单个参数$\langle k\rangle$决定。这样的指数分布的p_k通常出现在网络生长时没有偏好连接机制的情况下（**4.6**节）。

如果知道每个国家的电网的$\langle k\rangle$值，我们就可以预测其对攻击的临界阈值f_c^{targ}。如图**7-26f**所示，对于$\langle k\rangle>1.5$的国家的电网，观测的f_c^{targ}和计算得出的f_c^{targ}很好地吻合（第1组）。然而，对于$\langle k\rangle<1.5$的电网（第2组），计算得出的f_c^{targ}低估了真实的f_c^{targ}，也就是说，这些国家的电网对于攻击的鲁棒性优于基于度分布的期望。如同我们接下来要介绍的，增强的鲁棒性和相应国家电网的可靠性相关。

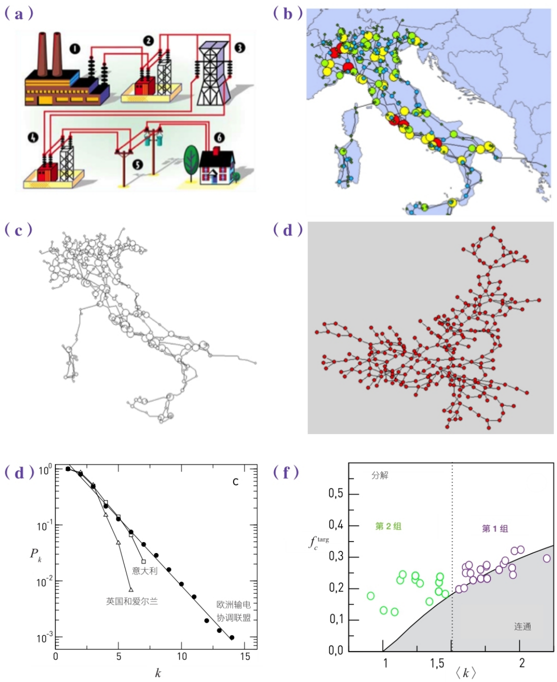


图7-26　电网

（a）　电网是一个复杂的基础设施，它包括（1）发电机、（2）切换单元、（3）高压输电网、（4）变压器、（5）低压输电线、（6）家用或商用消费者。在研究电网背后的网络时，我们略去了许多细节。

（b）～（d）　这三幅图描述的是，意大利电网及其发电和用电的具体情况。如果我们从电网中去掉具体细节，会得到图c中所示的空间网络。如果空间的信息也被移除，我们就得到了图d中所示的网络，这是一个典型的网络层面的研究对象。

（e）欧洲电网的度分布的互补累积分布P_k。该图显示的内容包括欧洲输电协调联盟的数据（UCTE）、意大利的数据、英国和爱尔兰的数据。这些国家的电网的度分布都符合公式7.19。

（f）图f描述的是，指数分布的不相关网络在受到攻击时的相空间(f_c^{targ}, $\langle k\rangle$)，其中f_c^{targ}是为了分解网络必须移除的枢纽节点占比。实线对应攻击的临界边界，当攻击处于这个边界下方时，网络能维持巨连通分支。该图还显示了33个欧盟国家的电网的$f_c^{\text{targ}}(\langle k\rangle)$的估计值，每个圆圈表示一个国家。图中显示了两类电网。对于$\langle k\rangle$ >1.5的国家（第1组），解析预测的f_c^{targ}与数值观测的值相符。对于$\langle k\rangle$ <1.5的国家（第2组），解析预测值低于数值观测的值。因此，第2组国家对攻击的鲁棒性有所增强。这意味着，在度序列相同的情况

下，这些真实网络比随机网络更具鲁棒性。[40]

我们通过分析从电网故障中收集和报告的以下量化数据来讨论鲁棒性和可靠性之间的关系：（1）未能供给的能量；（2）总共损失的电力；（3）平均中断的供电时间，即每年中断供电的分钟数。定量结果表明，真实和理论f_c^{targ}值相符的第1组网络（其网络规模占欧洲整体的2/3）与第2组中的网络承载了几乎同样多的电量。然而，第1组的平均中断时间是第2组的5倍，所记录的电力损失是第2组的2倍，未能传输的能量是第2组的4倍。因此，第1组的电网明显比第2组的电网更脆弱。这一结果证明了拓扑更鲁棒的网络其可靠性也更强。同时，这一发现是违反直觉的：人们通常期望更密集的网络会更具有鲁棒性。然而，我们发现，稀疏的电网却显示了增强的鲁棒性。

总而言之，为了提高复杂系统的鲁棒性，我们必须更好地理解网络的拓扑结构。我们可以设计网络拓扑结构来增强鲁棒性，让网络对随机故障和攻击都具有鲁棒性，也可以通过干预措施限制级联故障的扩散来增强鲁棒性。

这些结果或许表明，我们应该重新设计互联网和电网的拓扑，以增强其鲁棒性[44]。如果有这样的机会，这确实是可以实现的。然而，这些基础设施的网络在经过几十年的逐步建设中，经历了前几章所描述的自组织增长过程。鉴于每一个节点和链接所花费的巨大成本，我们不太可能有机会重建。

7.8 小结：阿喀琉斯之踵

“9·11事件”的主谋对攻击目标的选择绝非随意，纽约的世界贸易中心、五角大楼、华盛顿特区的白宫（预定目标）分别是美国经济、军事和政治的中心节点[45]。然而，尽管这次人类的悲剧远远超过了越南战争以来美国经历的任何其他事件，但这次攻击并未摧毁整个网络。的确，攻击者为美国发动新的战争提供了借口，比如伊拉克战争和阿富汗战争，并引发了一系列的级联事件，其影响比“9·11”事件本

身造成的破坏大得多。然而，所有的网络，无论是经济、军事，还是政治网络，都继续发挥了功效。因此，我们可以把“9·11”事件视为鲁棒性和网络韧性的故事（边栏7.5）。本章揭示了这种鲁棒性的来源：真正的网络的枢纽节点有一整套结构，取出其中任何一个都不足以摧毁整个网络。

边栏7.5

鲁棒性、韧性、冗余

冗余和韧性是与鲁棒性密切相关的两个概念，明确它们之间的差异对我们来说是有益处的。

鲁棒性

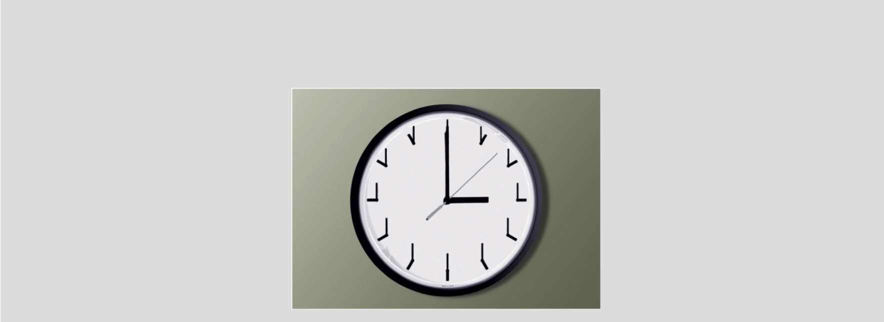
如果一个系统在内部和外部发生错误时仍能保持其基本功能，那么这个系统具有鲁棒性。在网络中，鲁棒性是指系统丢失一些节点和链接后仍能执行其基本功能的能力。

韧性

如果一个系统可以改变自身的运行模式来适应内部和外部的错误，且仍能执行其功能，那么这个系统就具有韧性。因此，韧性是一个要求系统改变核心活动的动力学属性。

冗余

系统中存在并行的组件和功能，且能在需要时相互替代丢失的组件和功能，我们称为冗余。网络中大部分的节点对之间存在多条独立路径，因此网络在信息导航方面有相当大的冗余。



真实网络体现的强大的鲁棒性对很多复杂系统来说是一个好消息。事实上，我们的细胞中发生过无数的错误，包括蛋白质错误折叠和转录因子延迟到位。然而，细胞网络的鲁棒性使细胞可以正常工作。网络鲁棒性也解释了为什么我们很少能注意到路由器故障对互联网的影响，以及为什么一个物种的消失不会立刻给环境带来灾难。

然而，这种拓扑鲁棒性也有弱点：受到攻击时是脆弱的。正如我们在这一章中所展示的，同时移除几个枢纽节点就可以分解任何网络。这对互联网来说是个坏消息，因为破坏者可以据此设计策略，来破坏这个重要的通信系统。对经济系统来说，这也是个坏消息，因为它表明，移除枢纽节点可以破坏整个经济，2009年的金融风暴就是活生生的例子。然而，它对药物设计来说是一个好消息，它表明，准确地理解细胞网络可以帮助我们开发药物，从而杀死不需要的细菌或癌细胞。

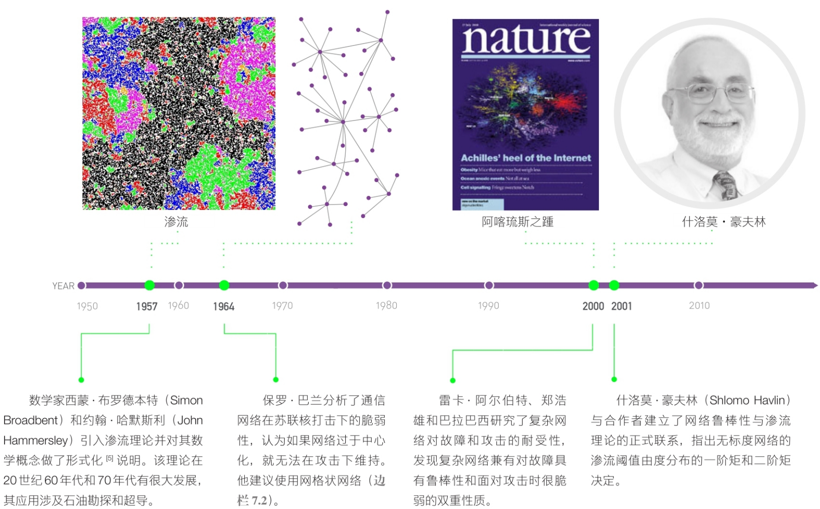


图7-27　从渗流到鲁棒性：一段简史

对网络鲁棒性的系统性研究从一篇在《自然》杂志上发表的一篇论文开始（图7-1）。这篇文章由雷卡·阿尔伯特、郑浩雄和巴拉巴西所写[1]，报告了无标度网络对随机故障的鲁棒性及面对攻击时的脆弱性。然而，对网络鲁棒性的解析依赖于渗流理论。在这种背景下，什洛莫·豪夫林和其合作者做出了重要的贡献。他们建立了鲁棒性与渗流理论的正式联系，并展示了无标度网络的渗流阈值由度分布的矩决定。豪夫林是一名来自以色列的统计物理学家，他对网络的研究有诸多贡献，其中包括发现真实网络的自相似性[46]和探索分层网络的鲁棒性[47]。

这一章的信息很简单：网络拓扑结构、鲁棒性和脆弱性是密不可分的。然而，每一个复杂的系统有它自己的“阿喀琉斯之踵”：系统背后的网络面对随机故障时具有鲁棒性，但同时在攻击下是脆弱的。

在考虑鲁棒性时，我们不能忽略这样的事实：大多数系统都有许多控制和反馈回路，帮助它们在遇到错误和故障时仍能维持运转。互联网协议被设计为“绕过故障寻路”，来引导信息流离开故障路由器；细胞有许多机制来去掉错误的蛋白质，并让功能错乱的基因停止运转。本章讨论了另一种增强鲁棒性的方式：系统背后的网络结构增强了系统的容错性。

无标度网络的鲁棒性不禁令人疑惑：增强的鲁棒性会不会是许多实际网络成为无标度网络的原因？也许真实系统正是为了满足其对鲁棒性的需要才演化出了无标度网络的结构。如果这个假设是正确的，那么我们将鲁棒性作为优化的准则，应该会得到一个无标度网络。然而，正如7.7节中所展示的，最具鲁棒性的网络是中心辐射拓扑结构。它的度分布是双峰的，而不是幂律的。这表明鲁棒性并不是推动真实网络演化的内在原因。相反，网络的无标度特性来源于网络的生长和偏好连接机制。虽然无标度网络恰好也有增强的鲁棒性，但我们所能设计出的最具鲁棒性的网络并不是它们。

边栏7.6

快速回顾：网络的鲁棒性

莫洛伊 - 里德准则：

存在巨连通分支的条件为 $\frac{\langle k^2 \rangle}{\langle k \rangle} > 2$

随机故障：

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$$

随机网络中：$f_c^{\mathrm{ER}} = 1 - \frac{1}{\langle k \rangle}$

增强的鲁棒性：$f_c > f_c^{\mathrm{ER}}$

攻击：

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{\min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right)$$

级联故障：

$$p(s) \sim s^{-\alpha}$$

$$\alpha = \begin{cases} \frac{3}{2}, & \gamma \geqslant 3 \\ \frac{\gamma}{\gamma - 1}, & 2 < \gamma < 3 \end{cases}$$

7.9 课后习题

习题1　随机故障：无标度网络以外的网络

为以下度分布的网络计算临界阈值 f_c ：

（1）带指数截断的幂律；

（2）对数正态分布；

（3）狄拉克函数分布（Delta distribution，所有节点的度相同）。

假设所有网络都是无限大的不相关网络。度分布的函数形式、一阶矩、二阶矩请参照表**3-2**，讨论所得的结果对网络鲁棒性的影响。

习题2　相关网络中的临界阈值

生成同配、异配、中性三个随机网络，各有10^4个节点，度分布服从度指数$\gamma=2.2$的幂律分布。使用**6.5**节中描述的布鲁内特和索科洛夫的算法来生成网络。使用计算机模拟来研究这三个网络对随机故障的鲁棒性，并比较三个网络的$P_{\infty}(f)/P_{\infty}(0)$比值。哪一个网络最具鲁棒性？并解释原因。

习题3　真实网络的故障

对表**3-1**中的每个网络计算分解网络所需要的故障节点数，假设网络都是不相关网络。

习题4　社交网络中的阴谋

在一个由版主创立的虚拟社区里，管理员打算用分而治之的策略把社会网络分裂成一些孤立的连通分支。作为反对者中的一员，你想破坏他们的计划。有传言说，管理员会将两类人逐出社区，一类人有许多朋友，另一类人的朋友往往彼此认识。作为反对者的代表，你将决定保护哪类人：是那些朋友圈里的人高度相连的人，还是那些有很多朋友的人？为了做这个决定，你在网络上模拟了两种攻击：（1）移除集聚系数很高的节点；（2）移除度很高的节点。对于这两种攻击，请你在以下两个网络上研究巨连通分支的规模随移除节点占比的变化函数：

（**1**）用配置模型（**3.8**节）生成一个有$N=10^4$个节点的网络，网络的度分布符合$\gamma=2.5$的幂律分布。

（**2**）用图**8-16**和进阶阅读**8.B**中讨论的层次模型生成一个有$N=10^4$个节点的网络。

集聚系数和度，哪一个是最敏感的拓扑信息？保护哪一类人可以

最大程度地减少攻击带来的损失？如果所有人的信息（包括集聚系数和度等）都不可见，情况会不会更好？为什么？

习题5　网络中的雪崩现象

用埃尔德什－雷尼模型$G(N, p)$生成一个随机网络，用配置模型生成一个无标度网络，两个网络都有$N=10^3$个节点，平均度都为$\langle k\rangle=2$。假设每个节点上都有一个桶，每个桶至多可以承载和该节点的度一样多的沙粒。模拟以下过程：

（1）在每一个时间段上，随机选取一个节点i，添加一粒沙。

（2）如果节点i上的沙粒数目达到或者超过了桶的容量，该节点会变得不稳定，并把它所承载的所有沙粒倾倒给它的邻居节点。

（3）如果倾倒后某些邻居节点上的桶也变得不稳定，这些节点同样会倾倒它们的沙粒。这样的过程一直持续到不再有不稳定的桶存在。我们把这个倾倒的序列过程称为雪崩，从最初的扰动（添加一粒沙）开始，所有变得不稳定的节点的数目总和就是这个雪崩的规模s。

重复（1）～（3）这个过程10^4次。假设在每一个时间段，占比10^{-4}的沙粒在倾倒过程中丢失，就不会出现网络中所有的桶都堆满沙粒的情况。研究雪崩规模的分布$P(s)$。

7.10 进阶阅读7.A 无标度网络中的渗流

为了理解无标度网络在接近阈值（公式7.7）时是如何分解的，我们需要先计算相应的临界指数γ_p、β_p和ν。计算结果表明，与随机网络相比，无标度性质使这些指数的值发生了系统性的变化（7.2节）。

我们从概率P_∞ 开始讨论。P_∞ 是一个随机选取的节点属于巨连通分支的概率。根据公式7.2 ，这个概率在接近p_c （在移除节点的语境下就是f_c ）时遵循幂律。通过计算可知，在一个无标度网络中，指数β_p 对于度指数γ 的依赖关系为[7]，[48]，[49]，[50]，[51]：

$$\beta_p = \begin{cases} \frac{1}{3-\gamma} & 2 < \gamma < 3 \\ \frac{1}{\gamma-3} & 3 < \gamma < 4 \\ 1 & \gamma > 4 \end{cases} \quad (7.20)$$

尽管在随机网络中（对应γ >4），指数β_p =1，但我们真正感兴趣的无标度网络大多数是：β_p >1。因此，与随机网络相比，无标度网络中的巨连通分支在临界阈值附近分解得更快。

在p_c 附近，描述平均连通分支规模的临界指数符合下面的公式[48]：

$$\gamma_p = \begin{cases} 1 & \gamma > 3 \\ -1 & 2 < \gamma < 3 \end{cases} \quad (7.21)$$

我们可能对γ_p 在γ <3时是负数感到惊讶。然而，我们知道在γ <3时，网络中总是存在一个巨连通分支。因此，公式7.1 中的发散现象在这种情况下并不存在。

对有任意度分布的随机连接网络来说，有限大小的簇规模分布符合以下公式[48]，[50]，[51]：

$$n_s \sim s^{-\tau} e^{-s/s^*} \quad (7.22)$$

在这里，n_s 是规模为s 的簇的数量，s^* 是临界簇规模。在临界点附近有：

$$s^* \sim |p - p_c|^{-\sigma} \quad (7.23)$$

临界指数为

$$\tau = \begin{cases} \frac{5}{2} & \gamma > 4 \\ \frac{2\gamma - 3}{\gamma - 2} & 2 < \gamma < 4 \end{cases} \qquad (7.24)$$

$$\sigma = \begin{cases} \frac{3-\gamma}{\gamma-2} & 2 < \gamma < 3 \\ \frac{\gamma-3}{\gamma-2} & 3 < \gamma < 4 \\ \frac{1}{2} & \gamma > 4 \end{cases} \qquad (7.25)$$

当$\gamma > 4$时，我们又一次回到了随机网络的值$\tau = 5/2$和$\sigma = 1/2$。

总而言之，描述无标度网络分解的临界指数依赖于度指数γ。在$3 < \gamma < 4$的区域内，这也是成立的，尽管这一阶段的渗流相变发生在阈值f_c介于0和1之间时。只有当$\gamma > 4$时，才能观测到无穷维渗流所表明的平均场特征，说明了随机网络在随机故障下的反应。

7.11
进阶阅读7.B
莫洛伊 - 里德准则

这一节的目的是推导莫洛伊 - 里德准则。该准则允许我们计算任意网络的渗流阈值[6]。如果存在巨连通分支，那么平均来说，每个属于巨连通分支的节点必须至少有两个邻居（图**7-8**）。因此，在巨连通分支中随机选择一个节点i，度k_i应该至少为2。用$P(k_i | i \leftrightarrow j)$来表示已知网络中一个度为$k_i$的节点连接到巨连通分支中的一个节点$j$时$k_i$的条件概率，这个条件概率使我们能得到节点$i$的度的期望为[51]：

$$\langle k_i | i \leftrightarrow j \rangle = \sum_{k_i} k_i P(k_i | i \leftrightarrow j) = 2 \qquad (7.26)$$

换言之，$\langle k_i | i \leftrightarrow j \rangle \geq 2$，这是节点$i$属于巨连通分支的条件。我们可以把公式**7.26**右侧的概率重写为：

$$P(k_i \mid i \leftrightarrow j) = \frac{P(k_i \mid i \leftrightarrow j)}{P(i \leftrightarrow j)} = \frac{P(i \leftrightarrow j \mid k_i)\, p(k_i)}{P(i \leftrightarrow j)} \quad (7.27)$$

其中最后一步我们用到了贝叶斯定理。对于一个度分布为p_k 的网络，当不存在度相关性时，我们有：

$$P(i \leftrightarrow j) = \frac{2L}{N(N-1)} = \frac{\langle k \rangle}{N-1},\ P(i \leftrightarrow j \mid k_i) = \frac{k_i}{N-1} \quad (7.28)$$

这表明节点i 可以从N -1个节点中以概率1/(N -1)任意选择一个连接，这个过程重复k_i 次。回到公式7.26 ，我们可以得到：

$$\sum_{k_i} k_i P(k_i \mid i \leftrightarrow j) = \sum_{k_i} k_i \frac{P(i \leftrightarrow j \mid k_i)\, p(k_i)}{P(i \leftrightarrow j)}$$
$$= \sum_{k_i} k_i \frac{k_i p(k_i)}{\langle k \rangle} = \frac{\sum_{k_i} k_i^2 p(k_i)}{\langle k \rangle} \quad (7.29)$$

于是我们得到了莫洛伊 - 里德准则即公式7.4 ，存在巨连通分支的条件为：

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} > 2 \quad (7.30)$$

7.12
进阶阅读7.C
随机故障下的临界阈值

本节中，我们来推导公式7.7 。这个公式提供了随机节点移除时的临界阈值[7]，[51]。如果随机移除占比f 的节点，会产生两个后果：

- 它改变了一些节点的度，这是因为之前与被移除节点相连的节点失去了它们的链接[$k \rightarrow k' \leq k$]。
- 它也因此改变了度分布，因为被移除节点的邻居的度改变了[$p_k \rightarrow p'_{k'}$]。

具体而言，如果我们随机移除占比f 的节点，一个度为k 的节点变成度为k'的节点的概率为：

$$\binom{k}{k'} f^{k-k'}(1-f)^{k'} \quad k' \leqslant k \tag{7.31}$$

公式7.31 中第一个与f 有关的被选中的节点失去了$k-k'$条链接，其中每条链接失去的概率都是f ；后一项余下的k'条链接没有受到节点移除的影响，其中每条链接未受影响的概率是$1-f$ 。

在原网络中，随机选取的节点度为k 的概率为p_k ；在新网络中，随机选取的节点度为k'的概率为：

$$p'_{k'} = \sum_{k=k'}^{\infty} p_k \binom{k}{k'} f^{k-k'}(1-f)^{k'} \tag{7.32}$$

假设基于原来的度分布已知$\langle k \rangle$和$\langle k^2 \rangle$ 。我们的目标是在移除占比为f 的节点后，计算新的度分布$p'_{k'}$ 的$\langle k \rangle$和$\langle k^2 \rangle$ 。为此，我们有：

$$\begin{aligned}\langle k' \rangle_f &= \sum_{k'=0}^{\infty} k' p'_{k'} \\ &= \sum_{k'=0}^{\infty} k' \sum_{k=k'}^{\infty} p_k \left(\frac{k!}{k'!(k-k')!} \right) f^{k-k'}(1-f)^{k'} \\ &= \sum_{k'=0}^{\infty} \sum_{k=k'}^{\infty} p_k \frac{k(k-1)!}{(k'-1)!(k-k')!} f^{k-k'}(1-f)^{k'-1}(1-f)\end{aligned} \tag{7.33}$$

上式的求和域是图7-28 所示的三角形。我们可以验证，如果我们按照下式同时改变求和顺序和求和极限，结果不变：

$$\sum_{k'=0}^{\infty} \sum_{k=k'}^{\infty} = \sum_{k=0}^{\infty} \sum_{k'=0}^{k} \tag{7.34}$$

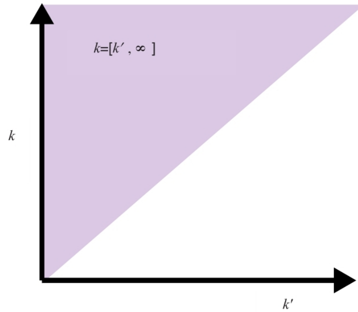


图7-28　积分域

在公式7.34　中，我们改变了积分的顺序，即改变了两次求和的先后顺序。之所以能够这样做，是因为两次都是在图中紫色三角形区域中求和。

因此，我们可以得到：

$$\begin{aligned}\langle k'\rangle_f &= \sum_{k=0}^{\infty} k' \sum_{k'=0}^{k} p_k \frac{k(k-1)!}{(k'-1)!(k-k')!} f^{k-k'}(1-f)^{k'-1}(1-f) \\ &= \sum_{k=0}^{\infty}(1-f)kp_k \sum_{k'=0}^{k} \frac{k(k-1)!}{(k'-1)!(k-k')!} f^{k-k'}(1-f)^{k'-1} \\ &= \sum_{k=0}^{\infty}(1-f)kp_k \sum_{k'=0}^{k} \binom{k-1}{k'-1} f^{k-k'}(1-f)^{k'-1} \\ &= \sum_{k=0}^{\infty}(1-f)kp_k \\ &= (1-f)\langle k\rangle\end{aligned} \quad (7.35)$$

这样，我们就把移除占比为f 的节点后的新网络的$\langle k'\rangle$ 与原网络的$\langle k\rangle$联系了起来。

对于$\langle k'^2\rangle$ ，我们也有相似的计算：

$$\begin{aligned}\langle k'^2\rangle_f &= \langle k'(k'-1)+k'\rangle_f \\ &= \langle k'(k'-1)\rangle_f + \langle k'\rangle_f \\ &= \sum_{k'=0}^{\infty} k'(k'-1)p'_{k'} + \langle k'\rangle_f\end{aligned} \quad (7.36)$$

再一次，我们改变求和顺序（图7-28 ），得到：

$$
\begin{aligned}
\left\langle k'(k'-1)\right\rangle_f &= \sum_{k'=0}^{\infty} k'(k'-1)p'_{k'} \\
&= \sum_{k'=0}^{\infty} k'(k'-1)\sum_{k=k'}^{\infty} p_k \binom{k}{k'} f^{k-k'}(1-f)^{k'} \\
&= \sum_{k=0}^{\infty} k'(k'-1)\sum_{k'=0}^{k} p_k \frac{k'(k'-1)}{k'!(k-k')!} f^{k-k'}(1-f)^{k'} \\
&= \sum_{k=0}^{\infty}\sum_{k=0}^{k} p_k \frac{k!}{(k'-2)!(k-k')!} f^{k-k'}(1-f)^{k'-2}(1-f)^2 \\
&= \sum_{k=0}^{\infty}(1-f)^2 k(k-1)p_k \sum_{k'=0}^{k} \frac{(k-2)!}{(k'-2)!(k-k')!} f^{k-k'}(1-f)^{k'-2} \\
&= \sum_{k=0}^{\infty}(1-f)^2 k(k-1)p_k \sum_{k'=0}^{k} \binom{k-2}{k'-2} f^{k-k'}(1-f)^{k'-2} \\
&= \sum_{k=0}^{\infty}(1-f)^2 k(k-1)p_k \\
&= (1-f)^2 \left\langle k(k-1)\right\rangle
\end{aligned}
\tag{7.37}
$$

因而我们有：

$$
\begin{aligned}
\left\langle k'^2\right\rangle_f &= \left\langle k'(k'-1)+k'\right\rangle_f \\
&= \left\langle k'(k'-1)\right\rangle_f + \left\langle k'\right\rangle_f \\
&= (1-f)^2\left\langle k(k-1)\right\rangle + (1-f)\left\langle k\right\rangle \\
&= (1-f)^2\left(\left\langle k^2\right\rangle - \left\langle k\right\rangle\right) + (1-f)\left\langle k\right\rangle \\
&= (1-f)^2\left\langle k^2\right\rangle - (1-f)^2\left\langle k\right\rangle + (1-f)\left\langle k\right\rangle \\
&= (1-f)^2\left\langle k^2\right\rangle - \left(-f^2+2f-1+1-f\right)\left\langle k\right\rangle \\
&= (1-f)^2\left\langle k^2\right\rangle + f(1-f)\left\langle k\right\rangle
\end{aligned}
\tag{7.38}
$$

这样我们就把移除了占比为f 的节点后新网络的$\left\langle k'^2\right\rangle$ 与原网络的$\left\langle k^2\right\rangle$ 联系了起来。我们把公式7.35 与7.36 写在一起：

$$
\left\langle k'\right\rangle_f = (1-f)\left\langle k\right\rangle \tag{7.39}
$$

$$
\left\langle k'^2\right\rangle_f = (1-f)^2\left\langle k^2\right\rangle + f(1-f)\left\langle k\right\rangle \tag{7.40}
$$

根据莫洛伊 - 里德准则（公式7.4 ），裂解阈值为：

$$
k = \frac{\left\langle k'^2\right\rangle_f}{\left\langle k'\right\rangle_f} = 2 \tag{7.41}
$$

把公式7.38 插入公式7.40 与7.41 中，我们最终得到公式7.7 ：

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1} \quad (7.42)$$

这是任意度分布p_k 的网络在随机节点移除后的裂解阈值。

7.13 进阶阅读7.D 分解有限规模的无标度网络

本节将推导在无标度网络上分解阈值与网络规模N 的关系（公式7.10 ）。我们从计算幂律分布的m 阶矩开始：

$$\langle k^m \rangle = (\gamma - 1)k_{\min}^{\gamma-1} \int_{k_{\min}}^{k_{\max}} k^{m-\gamma} dk$$

$$= \frac{(\gamma - 1)}{(m - \gamma + 1)} k_{\min}^{\gamma-1} \left[k^{m-\gamma+1} \right]_{k_{\min}}^{k_{\max}} \quad (7.43)$$

使用下列（公式3.18 ）：

$$k_{\max} = k_{\min} N^{\frac{1}{\gamma-1}} \quad (7.44)$$

我们有：

$$\langle k^m \rangle = \frac{(\gamma - 1)}{(m - \gamma + 1)} k_{\min}^{\gamma-1} \left[k_{\max}^{m-\gamma+1} - k_{\min}^{m-\gamma+1} \right] \quad (7.45)$$

为了计算f_c ，需要计算下面这个比例：

$$k = \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{(2 - \gamma)}{(3 - \gamma)} \frac{k_{\max}^{3-\gamma} - k_{\min}^{3-\gamma}}{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}} \quad (7.46)$$

在N 很大（$k_{\max}$ 自然也很大）的情况下，根据γ 的值可以化简为：

$$k = \frac{\langle k^2 \rangle}{\langle k \rangle} = \left| \frac{(2-\gamma)}{(3-\gamma)} \right| \begin{cases} k_{\min} & \gamma > 3 \\ k_{\max}^{3-\gamma} k_{\min}^{\gamma-2} & 3 > \gamma > 2 \\ k_{\max} & 2 > \gamma > 1 \end{cases} \quad (7.47)$$

根据公式7.7 给出的分解阈值：

$$f_c = 1 - \frac{1}{\kappa - 1} \quad (7.48)$$

其中k 由公式7.46 给出。把公式7.43 代入公式7.42 和公式7.47 中，我们得到：

$$f_c \approx 1 - \frac{C}{N^{\frac{3-\gamma}{\gamma-1}}} \quad (7.49)$$

也就是公式7.10 。

7.14 进阶阅读7.E 真实网络对攻击和故障的耐受性

在这一节中，我们将探讨表3-1 和公式7.2 中讨论的10个参照网络的攻击和故障曲线，如图7-29 所示。图中展示了几种不同的模式，证实了本章讨论的结果：

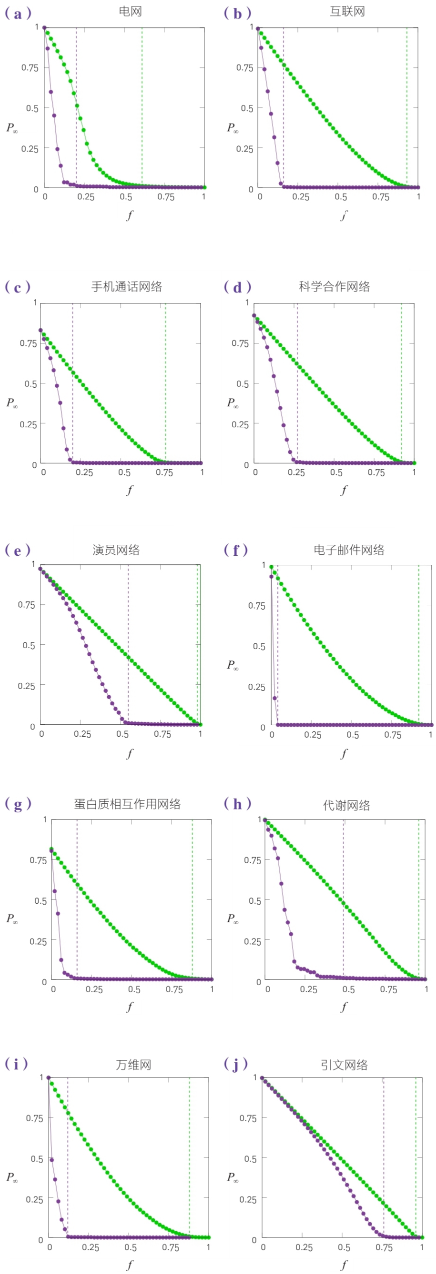


图7-29　故障与攻击曲线

故障（绿色）和攻击（紫色）曲线，对应表**3-1** 所示的10个参照网络。绿色竖线对

应对于随机故障的阈值f_c^{rand} 的估计值，紫色竖线对应攻击的阈值f_c^{targ} 的估计值。f_c 的估计值对应网络巨连通分支的占比首次低于其原始规模的1%的时候。在大多数系统中，这样的估计方法可以为f_c 给出很好的近似。唯一的例外是代谢网络，它的真实f_c^{targ} <0.25，但一个小簇的存在导致对f_c^{targ} 的估计值为f_c^{targ} ≈ 0.5。

- 对于所有的网络，故障和攻击曲线都是不同的。这正是我们讨论的阿喀琉斯之踵效应（**7.8**节　）：真实网络对随机故障具有鲁棒性，但面对攻击时又很脆弱。
- 故障和攻击曲线差异的程度取决于每个网络的平均度及度的异质性。例如，引文网络和演员网络的攻击阈值f_c 分别在0.5和0.75附近，这是相当大的值。因为这些网络非常密集，平均度为$\langle k \rangle$ $=20.8$（引文网络）和$\langle k \rangle$ $=83.7$（演员网络）。因此，这些网络在很大一部分枢纽节点被移除的时候，仍然可以保持连接。

7.15 进阶阅读7.F 攻击阈值

本节的目的是推导公式**7.12**　，即无标度网络的攻击阈值。我们的目标是计算不相关的无标度网络的f_c 。这个无标度网络由配置模型生成，度分布为$p_k = c \cdot k^{-\gamma}$，其中$k = k_{min}, \ldots, k_{max}$，且$c \approx (\gamma - 1) / (k_{min}^{-\gamma+1} - k_{max}^{-\gamma+1})$。

我们在网络中以度递减的顺序移除占比为f 的节点（即枢纽节点移除），会产生两个影响[9]，[51]：

（**1**）网络中的最大的度由k_{max} 变为k'_{max}。

（**2**）　被移除的枢纽节点的链接也被移除了，改变了余下网络中的度分布。

余下的网络仍然是不相关的。因此，我们可以使用莫洛伊－里德准则来判断巨连通分支是否存在。

先来考虑（1）带来的影响。新的最大度k'_{max}，由下列分式给出：

$$f=\int_{k'_{min}}^{k_{max}}p_k dk=\frac{\gamma-1}{\gamma-1}\frac{k'^{-\gamma+1}_{max}-k^{-\gamma+1}_{max}}{k^{-\gamma+1}_{min}-k^{-\gamma+1}_{max}} \quad (7.50)$$

如果我们假定$k_{max} \gg k'_{max}$ 且$k_{max} \gg k_{min}$（对于自然截断的大规模无标度网络这是成立的），可以忽略与k_{max}有关的项，得到：

$$f=\left(\frac{k'_{max}}{k_{min}}\right)^{-\gamma+1} \quad (7.51)$$

并进一步得出：

$$k'_{max}=k_{min}f^{\frac{1}{1-\gamma}} \quad (7.52)$$

公式7.52 给出了在移除占比为f 的枢纽节点后，网络中新的最大度的值。

下面，我们考虑（2），枢纽节点移除对度分布的影响为$p_k \to p'_k$。在不存在度相关性时，我们假定被移除的枢纽节点的链接是随机连到网络中其他节点的。因此，可以计算当f 比例的枢纽节点被移除后，一起被“随机”移除的链接在网络中的比例$\tilde{f}$：

$$\tilde{f}=\frac{\int_{k'_{max}}^{k_{max}}kp_k dk}{\langle k\rangle}=\frac{1}{\langle k\rangle}c\int_{k'_{max}}^{k_{max}}k^{-\gamma+1}dk$$

$$=\frac{1}{\langle k\rangle}\frac{1-\gamma}{2-\gamma}\frac{k'^{-\gamma+2}_{max}-k^{-\gamma+2}_{max}}{k^{-\gamma+1}_{min}-k^{-\gamma+2}_{max}} \quad (7.53)$$

再次忽略k_{max}这一项，并考虑$\langle k\rangle \approx \frac{\gamma-1}{\gamma-2}k_{min}$，我们得到：

$$\tilde{f}=\left(\frac{k'_{max}}{k_{min}}\right)^{-\gamma+2} \quad (7.54)$$

利用公式7.51 得到：

$$\tilde{f} = f^{\frac{2-\gamma}{1-\gamma}} \quad (7.55)$$

当$\gamma \to 2$时，我们有$\tilde{f} \to 1$，这意味着移除一小部分枢纽节点就移除了所有的链接，有可能已经摧毁了网络。这与第3章的发现一致：当$\gamma = 2$时枢纽节点主宰了整个网络。

一般情况下，余下网络的度分布为：

$$p'_{k'} = \sum_{k=k_{min}}^{k'_{max}} \binom{k}{k'} \tilde{f}^{k-k'} (1-\tilde{f})^{k'} p_k \quad (7.56)$$

注意，我们在进阶阅读7.C 中已经推导出了这个度分布7.32 。我们可以借用其计算方法来推导随机节点移除的情况。具体来说，根据公式7.45 ，可以用k_{min}和k'_{max} 来计算无标度网络中的k ：

$$\kappa = \frac{2-\gamma}{3-\gamma} \frac{k'^{3-\gamma}_{max} - k^{3-\gamma}_{max}}{k'^{2-\gamma}_{max} - k^{2-\gamma}_{max}} \quad (7.57)$$

将公式7.52 代入上式可得：

$$\begin{aligned} \kappa &= \frac{2-\gamma}{3-\gamma} \frac{k^{3-\gamma}_{min} f^{(3-\gamma)/(1-\gamma)} - k^{3-\gamma}_{min}}{k^{2-\gamma}_{min} f^{(2-\gamma)/(1-\gamma)} - k^{2-\gamma}_{min}} \\ &= \frac{2-\gamma}{3-\gamma} k_{min} \frac{f^{(3-\gamma)/(1-\gamma)} - 1}{f^{(2-\gamma)/(1-\gamma)} - 1} \end{aligned} \quad (7.58)$$

简单变形后得到：

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right) \quad (7.59)$$

也就是公式7.12 。

7.16
进阶阅读7.G
最优度分布

在这一节中，我们要推导在第7.7节 中讨论的双峰度分布。这种网络的拓扑结构可以同时优化网络面对攻击和故障时的鲁棒性 [37] 。正如我们在公式7.17 中所做的，双峰分布由两个δ 函数组成：

$$p_k = (1-r)\delta(k-k_{\min}) + r\delta(k-k_{\max}) \qquad (7.62)$$

我们先来计算当$\langle k\rangle$ 固定时，总临界阈值f^{tot} 关于r 和$k_{\max}$ 的函数。为了得到f^{rand} 和f^{targ} 的解析表达，需要计算双峰分布（公式7.62 ）的矩：

$$\langle k\rangle = (1-r)k_{\min} + rk_{\max}$$
$$\langle k^2\rangle = (1-r)k^2_{\min} + rk^2_{\max} = \frac{(\langle k\rangle - rk_{\max})^2}{1-r} + rk^2_{\max} \qquad (7.63)$$

把上式代入公式7.7 可得：

$$f_c^{\text{rand}} = \frac{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} - 2(1-r)\langle k\rangle + rk^2_{\max}}{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} - (1-r)\langle k\rangle + rk^2_{\max}} \qquad (7.64)$$

为了确定针对性攻击的阈值，我们要考虑到只有两种节点：占比为r 的度为$k_{\max}$ 的节点，还有占比为(1-r)的度为$k_{\min}$ 的节点。因此，枢纽节点移除要么会移除所有的枢纽节点，如同情况（1），要么会只移除部分枢纽节点，如同情况（2）：

（1） $f_c^{\text{targ}} > r$ 。在这种情况下，所有的枢纽节点都被移除。因此，针对性攻击后余下的节点的度均为$k_{\min}$ ，我们可以得出：

$$f_c^{\text{targ}} = r + \frac{1-r}{\langle k\rangle - rk_{\max}}\left\{\langle k\rangle \frac{\langle k\rangle - rk_{\max} - 2(1-r)}{\langle k\rangle - rk_{\max} - (1-r)} - rk_{\max}\right\} \qquad (7.65)$$

（2） $f_c^{\text{targ}} < r$。在这种情况下，被移除节点都是大度节点，并且还有一些大度节点$k_{\max}$ 剩下。于是我们得出：

$$f_c^{\text{targ}} = \frac{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} + rk^2_{\max} - 2(1-r)\langle k\rangle}{k_{\max}(k_{\max}-1)(1-r)} \qquad (7.66)$$

利用公式7.64 ~7.66 这些阈值，我们可以计算总阈值f^{tot}（公式7.16 ）。为了得到k_{max}最优值关于r 的函数的表达式，我们取f^{tot}最大化时的k 值。利用公式7.64 和7.66 ，我们发现在r 较小的情况下，k_{max}的最优值可被近似为：

$$k_{max} \sim \left\{ \frac{2\langle k\rangle^2 (\langle k\rangle - 1)^2}{2\langle k\rangle - 1} \right\}^{1/3} r^{-2/3} = Ar^{-2/3} \quad (7.67)$$

利用这一结果和公式7.16 ，在r 较小的情况下，我们得出：

$$f_c^{tot} = 2 - \frac{1}{\langle k\rangle - 1} - \frac{3\langle k\rangle}{A^2} r^{1/3} + O(r^{2/3}) \quad (7.68)$$

因此，f^{tot} 在r 趋近于零时达到理论最大值。对于有N 个节点的网络，$r = 1/N$ （满足至少存在一个节点的度为k_{max} 的最小r 值）时，f^{tot} 取得最大值。在这种情况下，决定中央枢纽节点的度k_{max} 的最优值的方程如下[37] ：

$$k_{max} = AN^{2/3} \quad (7.69)$$

这里的A 如公式7.67 所定义。

7.17 参考文献

[1] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature* , 406: 378, 2000.

[2] D. Stauffer and A. Aharony. *Introduction to Percolation Theory* . Taylor and Francis. London, 1994.

[3] A. Bunde and S. Havlin. *Fractals and Disordered Systems* .Springer, 1996.

[4] B. Bollobás and O. Riordan. *Percolation* . Cambridge

University Press. Cambridge, 2006.

[5] S. Broadbent and J. Hammersley. Percolation processes I.Crystals and mazes. *Proceedings of the Cambridge Philosophical Society* , 53: 629, 1957.

[6] M. Molloy and B. Reed. A criticial point for random graphs with a given degree sequence. *Random Structures and Algorithms* , 6: 161, 1995.

[7] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. Resilience of the Internet to random breakdowns. *Physical Review Letters* , 85: 4626, 2000.

[8] D. S. Callaway, M. E. J. Newman, S. H. Strogatz. and D. J.Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters* , 85: 5468–5471, 2000.

[9] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. Breakdown of the Internet under intentional attack. *Physical Review Letters* , 86: 3682, 2001.

[10] B. Bollobás and O. Riordan. Robustness and Vulnerability of Scale-Free Random Graphs. *Internet Mathematics* , 1: 1-35, 2003.

[11] P. Baran. Introduction to Distributed Communications Networks. *Rand Corporation Memorandum* , RM-3420-PR, 1964.

[12] D.N. Kosterev, C.W. Taylor and W.A. Mittlestadt. Model Validation of the August 10, 1996 WSCC System Outage. *IEEE Transactions on Power Systems* 14: 967-979, 1999.

[13] C. Labovitz, A. Ahuja and F. Jahasian. Experimental Study of Internet Stability and Wide-Area Backbone Failures. Proceedings of IEEE FTCS, Madison, WI, 1999.

[14] A. G. Haldane and R. M. May. Systemic risk in banking ecosystems. *Nature* , 469: 351-355, 2011.

[15] T. Roukny, H. Bersini, H. Pirotte, G. Caldarelli and S.Battiston. Default Cascades in Complex Networks: Topology and Systemic Risk. *Scientific Reports* , 3: 2759, 2013.

[16] G. Tedeschi, A. Mazloumian, M. Gallegati, and D.Helbing. Bankruptcy cascades in interbank markets. *PLOS One* , 7:e52749, 2012.

[17] D. Helbing. Globally networked risks and how to respond. *Nature* , 497: 51-59, 2013.

[18] I. Dobson, B. A. Carreras, V. E. Lynch and D. E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *CHAOS* , 17: 026103, 2007.

[19] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 65-74, 2011.

[20] Y. Y. Kagan. Accuracy of modern global earthquake catalogs. *Phys* . *Earth Planet. Inter* ., 135: 173, 2003.

[21] M. Nagarajan, H. Purohit, and A. P. Sheth. A Qualitative Examination of Topical Tweet and Retweet Practices. *ICWSM* , 295-298, 2010.

[22] P. Fleurquin, J.J. Ramasco and V.M. Eguiluz. Systemic delay propagation in the US airport network. *Scientific Reports* , 3: 1159, 2013.

[23] B. K. Ellis, J. A. Stanford, D. Goodman, C. P. Stafford,D.L. Gustafson, D. A. Beauchamp, D. W. Chess, J. A. Craft, M. A.Deleray, and B. S. Hansen. Long-term effects of a trophic cascade in a large lake ecosystem. *PNAS* , 108: 1070, 2011.

[24] V. R. Sole, M. M. Jose. Complexity and fragility in

ecological networks. *Proc. R. Soc. Lond. B* , 268: 2039, 2001.

[25] F. Jordán, I. Scheuring and G. Vida. Species Positions and Extinction Dynamics in Simple Food Webs. *Journal of Theoretical Biology* , 215: 441-448, 2002.

[26] S.L. Pimm and P. Raven. Biodiversity: Extinction by numbers. *Nature* , 403: 843, 2000.

[27] World Economic Forum, *Building Resilience in Supply Chains* . World Economic Forum, 2013.

[28] Joint Economic Committee of US Congress. Your flight has been delayed again: Flight delays cost passengers, airlines and the U.S. economy billions. Available at http://www.jec.senate.gov, May 22. 2008.

[29] I. Dobson, A. Carreras, and D.E. Newman. A loading dependent model of probabilistic cascading failure. *Probability in the Engineering and Informational Sciences* , 19: 15, 2005.

[30] D.J. Watts. A simple model of global cascades on random networks. *PNAS* , 99: 5766, 2002.

[31] K.-I. Goh, D.-S. Lee, B. Kahng, and D. Kim. Sandpile on scale-free networks. *Physical Review Letters* , 91: 148701, 2003.

[32] D.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim. Sandpile avalanche dynamics on scale-free networks. *Physica A* , 338: 84, 2004.

[33] M. Ding and W. Yang. Distribution of the first return time in fractional Brownian motion and its application to the study of onoff intermittency. *Physical Review Letters* , 52: 207-213, 1995.

[34] A. E. Motter and Y.-C. Lai. Cascade-based attacks on complex networks. *Physical Review Letters* , 66: 065102, 2002.

[35] Z. Kong and E. M. Yeh. Resilience to Degree-Dependent

and Cascading Node Failures in Random Geometric Networks. *IEEE Transactions on Information Theory* , 56: 5533, 2010.

[36] G. Paul, S. Sreenivas, and H. E. Stanley. Resilience of complex networks to random breakdown. *Physical Review Letters* , 72: 056130, 2005.

[37] G. Paul, T. Tanizawa, S. Havlin, and H. E. Stanley. Optimization of robustness of complex networks. *European Physical Journal* B, 38: 187–191, 2004.

[38] A.X.C.N. Valente, A. Sarkar, and H. A. Stone. Two-peak and three-peak optimal complex networks. *Physical Review Letters* , 92: 118702, 2004.

[39] T. Tanizawa, G. Paul, R. Cohen, S. Havlin, and H. E.Stanley. Optimization of network robustness to waves of targeted and random attacks. *Physical Review Letters* , 71:047101, 2005.

[40] R.V. Sole, M. Rosas-Casals, B. Corominas-Murtra, and S. Valverde. Robustness of the European power grids under intentional attack. *Physical Review Letters* , 7 7: 026102, 2008.

[41] R. Albert, I. Albert, and G.L. Nakarado. Structural Vulnerability of the North American Power Grid. *Physical Review Letters* , 69: 025103 R, 2004.

[42] A.E. Motter. Cascade control and defense in complex networks. *Physical Review Letters* , 93: 098701, 2004.

[43] A. Motter, N. Gulbahce, E. Almaas, and A.-L. Barabási. Predicting synthetic rescues in metabolic networks. *Molecular Systems Biology* , 4: 1-10, 2008.

[44] C.M. Schneider, N. Yazdani, N.A.M. Araújo, S. Havlin and H.J. Herrmann. Towards designing robust coupled networks. *Scientific Reports* , 3: 1969, 2013.

[45] A.-L. Barabási. *Linked: The New Science of Networks* .Plume,

New York, 2002.

[46] C.M. Song, S. Havlin, and H.A Makse. Self-similarity of complex networks. *Nature* , 433: 392, 2005.

[47] S.V. Buldyrev, R. Parshani, G. Paul, H.E. Stanley and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature* , 464: 08932, 2010.

[48] R. Cohen, D. ben-Avraham and S. Havlin. Percolation critical exponents in scale-free networks. *Physical Review Letters* , 66: 036113, 2002.

[49] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Anomalous percolation properties of growing networks. *Physical Review Letters* , 64: 066110, 2001.

[50] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review Letters* , 64: 026118, 2001.

第8章
社区



图**8-0**　艺术与网络：**K-Mode**，年轻模式

艺术集K-Mode是发布各个概念下新奇趋势的报告。此图是《年轻模式：关于自由的报告》中的一页，讨论了“社区”的本义与含义之间的微妙转变。[1] 这也是本章的主题。

本章索引

8.1　本章概览 / 399

8.2　社区基础 / 402

8.3　层次聚类 / 408

8.4　模块度 / 417

8.5　重叠社区 / 425

8.6　检验社区 / 431

8.7　刻画社区特征 / 437

8.8　小结 / 441

8.9　课后习题 / 445

8.10　进阶阅读8.A
层次模块度 / 447

8.11　进阶阅读8.B
模块度 / 452

8.12　进阶阅读8.C
社区识别的快速算法 / 454

8.13　进阶阅读8.D
团渗流的阈值 / 459

8.14　参考文献 / 461

本章致谢

马顿·波什福伊　**MÁRTON PÓSFAI**

妮科尔·萨迈　**NICOLE SAMAY**

萨拉·莫里森　**SARAH MORRISON**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

罗伯塔·西纳特拉　**ROBERTA SINATRA**

8.1
本章概览

比利时似乎是典型的双文化社会：59%的公民是说荷兰语的弗拉芒人，40%的公民是说法语的瓦隆人。随着全世界的多民族国家纷纷分裂，我们不禁要问：比利时如何促进了两个民族自1830年以来的和平共处？这是不是一个无论弗拉芒人还是瓦隆人都同样紧密相连的社会？抑或在这个国家里存在彼此联系不多的两个小国家？

文森特·布隆代尔（Vincent Blondel）和他的学生在2007年给出了答案。他们发明了一种算法来识别这个国家的社区结构。他们从手机通话网络开始，把频繁打电话的个体放在一起[2]。算法表明，比利时的社交网络分裂为两个大的社群，每个社群内部的个体极少与另一个社群中的个体通话（图**8-1**）。他们给每个个体节点标注了其所使用的语言后，这一分裂的原因便显而易见：其中一个群体（cluster）几乎只包括说法语的人，另一个群体中则都是说荷兰语的人。

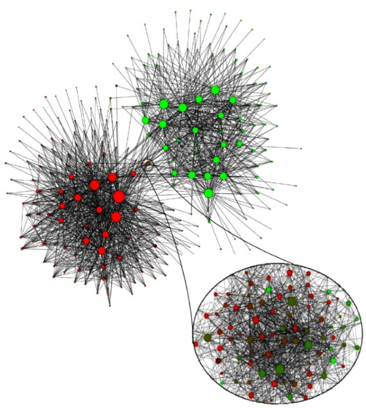


图**8-1** 比利时的社区

此图描述的是从比利时最大的手机公司的用户通话模式中提取出的社区。网络中有大约200万手机用户。图中的节点表示社区，节点的大小与社区人数成正比。社区的颜色表示用户所说的语言，红色表示法语，绿色表示荷兰语。图中只显示了多于100人的社区。连接两个主要社群的社区由多个较小社区组成，这些社区来自具有多元文化

的首都布鲁塞尔，那里的语言分化并不明显。[2]

在网络科学中，如果一组节点内部链接紧密，外部链接稀疏，那么我们称这组节点为一个社区（community）。接下来我们讨论社区发挥重要作用的两个场景，以便直观地理解社区的意义。

社交网络

研究人员在几十年前就已注意到，很容易在社交网络中找到社区[3-7]。事实上，一个公司的雇员更可能与本公司的同事交流，而不是与其他公司的雇员交流[3]。因此工作场所是社交网络中紧密相连的社区。同时，社区也可以表示朋友圈子、有相同爱好的人，或居住在同一个小区的人。

在关于社区的研究中，一个叫作扎卡里空手道俱乐部（Zachary's Karate Club）的社交网络受到了大量关注（图**8-2**）[7]。该网络包含了一家空手道俱乐部中34个成员之间的连接关系。由于这个俱乐部规模很小，每个成员都相互认识，但这种认识并不代表成员之间真正的社交关系。为了挖掘成员之间的真正关系，社会学家韦恩·扎卡里（Wayne Zachary），通过观察成员是否在俱乐部以外有定期交流，最终确定了78对这样的成员关系（图**8-2a**）。

研究者对这一数据集的兴趣源自一个特殊事件：当时俱乐部主席和教练不和，将俱乐部一分为二。大约一半成员追随教练，另一半追随主席。这一分裂行为揭示了网络潜在的社区结构（图**8-2a**）。今天对社区识别算法的测试通常都基于这一数据集，检验一个算法是否能从网络分裂前的结构推断出这两个社区。

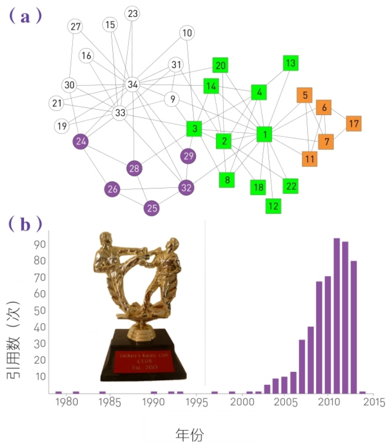


图**8-2**　扎卡里空手道俱乐部

（**a**）　此图描述的是扎卡里空手道俱乐部里34个成员之间的联系。链接表示俱乐部成员在俱乐部外的交流。圆圈和方块表示俱乐部一分为二后出现的两部分。颜色表示优化模块度系数*M* 的算法预测的最佳社区划分（**8.4**节　）。社区边界与实际的分裂很接近：白色和紫色社区是一部分，绿色和橙色社区是另一部分。[8]

（**b**）扎卡里空手道俱乐部论文[7] 的引用历史反映了网络科学中社区识别的历史。事实上，在2002年米歇尔·格文和马克·纽曼将扎卡里的论文作为社区识别的基准[9] 之前，几乎没有人对扎卡里的论文感兴趣。从那时起，这篇论文的引用数爆炸式增长，让人想起了在无标度网络发现后，埃尔德什和雷尼的化文引用数激增（图**2-15**　）。

扎卡里空手道俱乐部网络在社区研究领域的频繁使用催生了"扎卡里空手道俱乐部"。该俱乐部有一个不成文的规则："在任何网络会议上，首个研究扎卡里的空手道俱乐部的科学家将被纳入俱乐部并获奖。"

因此这个奖并非基于个人能力，而是基于科学家的简单参与行为。然而，获奖者都是杰出的网络科学家。图中为"扎卡里空手道俱乐部"奖杯，由最近的一位获奖者持有。感谢玛丽安·伯格纳（Marián Boguñá）提供照片。

生物网络

在理解特定生物功能如何在细胞网络中编码时，社区扮演了相当重要的角色。在李·哈特维尔（Lee　Hartwell）获得诺贝尔生理学或医学奖之前两年，他认为生物学研究的重点必须往前更进一步，从单个基因研究转变到探索分子群如何构成功能性模块，以实现特定的细胞

功能[10]。劳沃斯（Ravasz）和合作者[11]首次尝试在代谢网络中系统地识别这类模块。为此他们提出了一个算法，该算法能够识别具有局部紧密社区的分子群（图**8-3**）。

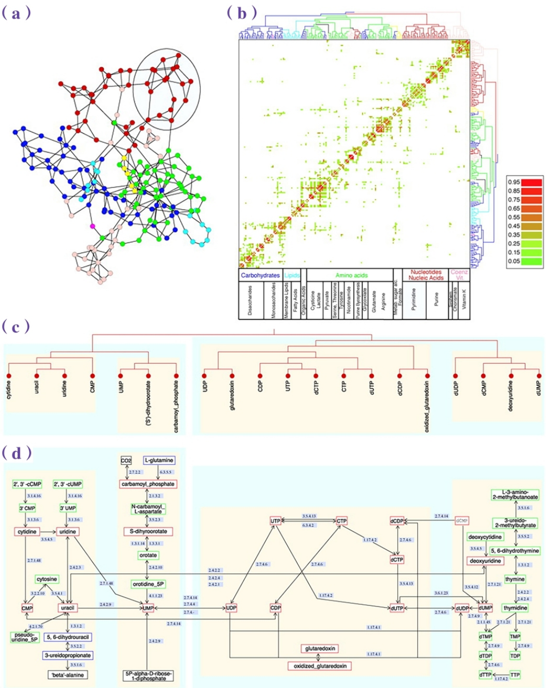


图**8-3**　代谢网络的社区

大肠杆菌的代谢为研究生物系统的社区结构提供了良好的基础[11]。

（**a**）劳沃斯算法识别的生物模块（社区）[11]（**8.3**节）。每个节点的颜色表示其主要从属的生化分类。可见不同的功能分类被分隔在网络的不同邻域中。高亮区域的节点属于嘧啶代谢，亦即算法发现的社区之一。

（**b**）大肠杆菌代谢的拓扑重叠矩阵和对应的树状图。从树状图中可以识别图a中所示的模块。每个分支的颜色反映了其中的分子的主要生化作用，例如碳水化合物（蓝色）、核苷酸和核酸代谢（红色）、脂代谢（青色）。

（**c**）在这两幅图中，树状图右侧的红色分支对应吡啶模块。

（**d**）嘧啶模块的具体代谢反应。反应周围的方框是劳沃斯算法发现的社区。[11]

在理解人类疾病时，社区也扮演了相当重要的角色。事实上，一

种疾病中涉及的多种蛋白质往往会相互作用[12]，[13]。基于这一发现，有研究者提出了疾病模块假说（disease module hypothesis）[14]，认为每种疾病都对应细胞网络中一个严格定义的邻域（neighborhood）。

上述示例展示了识别社区结构的多种动机。社区存在的根源是谁与谁相连，不能仅用度分布来解释。因此，要识别社区，我们必须检查网络的具体连接模式。上述示例引出了本章的一些假设：

H1：基础假设

网络的社区结构仅由其连接模式决定。

根据该假设，网络的社区结构可以通过检查网络邻接矩阵A_{ij}来发现。

本章的目的是介绍理解和识别一个复杂网络的社区结构所需的概念。我们将讨论如何定义社区，探索不同的社区性质，介绍基于不同原则的一系列社区识别算法。

8.2 社区基础

社区究竟意味着什么？一个网络中有多少个社区？一个网络能用多少种不同的方式进行社区划分？在这一节中，我们将讨论这些经常在社区识别中出现的问题。

定义社区

我们对社区的感性认知来自第二个假设（图**8-4**）：

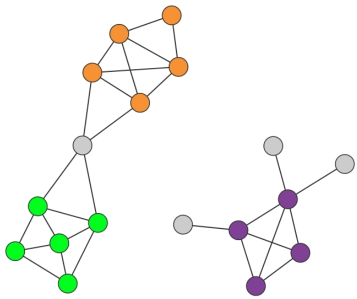


图8-4　连通性和密度假设

社区是网络中局部紧密连接的子图。这一断言依赖于两个假设：

连通性假设

每个社区对应一个连通子图，例如由橙色、绿色或紫色节点分别构成的子图。因此，如果网络中有两个不相连的分支，则每个社区只能存在于一个分支内。同时，一个连通分支内的两个互不相连的子图不能组成社区。因此，橙色和绿色节点只能构成不同的社区。

密度假设

社区中的节点更倾向于连接到同一社区内的其他节点，而非其他社区的节点。橙色、绿色和紫色节点都满足这一判断。

H2：连通性和密度假设

社区是网络中局部紧密连接的子图。

换言之，一个社区中的所有成员都可以通过同一社区中的其他成员到达（连通性，connectedness）。同时，我们希望一个社区内部的节点连接到同一社区中其他节点的概率高于连接到不在同一社区的节点的概率（密度，density）。虽然这一假设很大程度上缩小了“什么可被视为社区”的范围，但并没有给出关于社区的唯一定义。事实上，如同我们接下来所讨论的，有多种关于社区的定义与H2保持一致。

最大团

关于社区结构的早期论文之一（发表于1949年）将一个所有成员

都相互认识的群体定义为社区[5]。用图论的术语来讲，社区是一个完全子图，或者说一个团。团自然满足H2：是连通子图且有最大的链接密度。然而，若将社区视为团，会有如下缺陷：

- 虽然网络中存在大量三角形，但更大的团很少出现。
- 要求社区是完全子图可能太严苛了，很多合理的社区都不满足这一条件。例如，图**8-2** 和图**8-3** 中的社区无一是完全子图。

强社区和弱社区

对上述严苛的定义稍做放松，我们来看看网络的一个具有N_c 个节点的连通子图C 。节点i 的内部度（internal degree）k_i^{int} 是i 与C 中其他节点的链接数。外部度（external degree）k_i^{ext} 是i 与网络中其他节点（不属于子图C 的节点）的链接数。若$k_i^{\text{ext}}=0$，则i 的每个邻居都在C 中，因此应将C 视为节点i 所属的社区。若$k_i^{\text{int}}=0$，则节点i 应该属于另一个社区。这一定义允许我们区分两种社区（图**8-5** ）：

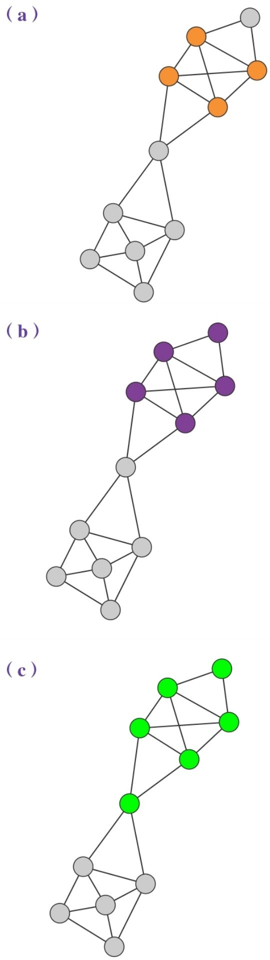


图8-5　定义社区

(a) 团

团对应完全子图。图中最高阶的团是橙色节点构成的四方形。网络中还有一些由3个节点构成的团，你能找到吗？

(b) 强社区

公式**8.1** 中定义的强社区是一个连通子图，其节点指向同一社区内其他节点的链接多于指向其他社区的节点的链接。紫色节点构成满足这一定义的强社区。图中还有一些强社区，你能再找到至少两个吗？

(c) 弱社区

公式**8.2** 中定义的弱社区是一个子图，其节点的总内部度高于总外部度。绿色节点

表示网络中多个可能的弱社区之一。

强社区

若C 中每个节点与社区内部节点的链接多于同社区外部节点的链接 [15] ， [16] ，则C 为强社区。具体而言，子图C 构成强社区需要每个节点$i \in C$ 满足：

$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C) \quad (8.1)$$

弱社区

若C 中所有节点的内部度之和高于所有节点的外部度之和 [16] ，则C 为弱社区。具体而言，子图C 构成弱社区需要满足：

$$\sum_{i \in C} k_i^{\text{int}}(C) > \sum_{i \in C} k_i^{\text{ext}}(C) \quad (8.2)$$

相对于强社区的要求，弱社区的定义放松了对节点的约束，允许一些节点不满足公式**8.1** 。换言之，不等式**8.2** 定义在整个社区上，而不是在单个节点上。

注意，每个团都是强社区，每个强社区都是弱社区。反过来则通常不成立（图**8-5** ）。

上述对社区的定义（团、强社区、弱社区）增进了我们对社区的理解，同时表明我们在社区定义上尚有自由发挥的空间。

社区数目

有多少种方式将网络中的节点组织成社区？若想回答这个问题，我们需要考虑最简单的社区识别问题，即图的二划分（graph bisection）问题：将一个网络分成两个不重叠的子图，使得两部分节点之间的链接数（被称为割集的大小）最小（边栏**8.1** ）。

边栏**8.1**

图划分

芯片设计师面临着一个非常复杂的问题：需要在芯片上放25亿个晶体管，且电线不能相交。为了简化这个问题，他们首先把集成电路的连接图划分为较小的子图，使子图之间的链接数量尽可能减少，然后分别处理集成电路的各个部分，最后把它们重新连起来。并行计算也遇到了类似的问题，如何把大型计算任务划分成子任务并分配给各个芯片。应该尽可能地减少处理器之间的通信，因为这类通信通常比较慢。

芯片设计师和软件工程师面临的问题在计算机科学中被称为图划分 [17] 。针对这一问题所设计的算法，例如广泛使用的Kerninghan-Lin算法（图**8-6** ），正是我们本章讨论的社区识别算法的前身。

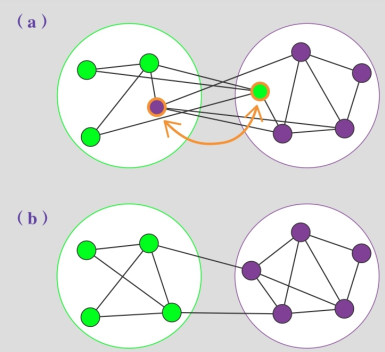


图**8-6　Kerninghan-Lin**算法

图划分的最佳算法是1970年提出的Kerninghan-Lin算法 [18] 。我们以图的二划分为例介绍这个算法：首先，随机把网络划分为预定大小的两组节点。接下来选择一对节点（i，j ），使得i 和j 分别在不同的组里，然后交换i 和j ，并记录交换后割集中的链接数。遍历所有节点对（i，j ）找到能最大限度减少割集中的链接数的节点对，例如图a中高亮的节点对。交换这一对节点，得到图b所示的划分。在该算法的某些实现过程中，如果所有节点对都无法减少割集中的链接数，则选择让割集中的链接数增长最少的节点对。

图划分和社区识别之间有重要的差别：图划分把网络分成预先指定数量的较小子图；与之相反，社区识别旨在揭示网络的固有社

区结构。因此，在大多数社区识别算法中，社区的数量和大小并非预先指定，而是需要通过检查网络连接模式来发现。

图划分

图的二划分问题可以这样求解：穷举所有把网络分成两部分的划分方式，选择其中割集最小的。为了确定这种蛮力的算法的开销，我们注意到，把一个具有N个节点的网络分成分别拥有N_1和N_2个节点的两部分，所有不同划分方式的数目为：

$$\frac{N!}{N_1!N_2!} \qquad (8.3)$$

使用斯特灵公式 $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ ，我们将公式**8.3** 改写为：

$$\frac{N!}{N_1!N_2!} \simeq \frac{\sqrt{2\pi N}\left(\frac{N}{e}\right)^N}{\sqrt{2\pi N_1}\,(N_1/e)^{N_1}\sqrt{2\pi N_2}\,(N_2/e)^{N_2}} \sim \frac{N^{N+1/2}}{N_1^{N_1+1/2}N_2^{N_2+1/2}} \qquad (8.4)$$

为了简化这一问题，我们只考虑把网络分成均匀的两部分，即$N_1=N_2=N/2$。在这种情况下，公式**8.4** 变成：

$$\frac{2^{N+1}}{\sqrt{N}} = e^{(N+1)\ln 2-\frac{1}{2}\ln N} \qquad (8.5)$$

表明图二划分的数目随着网络大小呈指数级增长。

为了直观展示公式**8.5** 的意义，考虑一个拥有10个节点的网络，将其划分成大小相同的两部分，即$N_1=N_2=5$。根公式**8.3** ，我们需要检查252种二划分方式来寻找最小割。假设我们的计算机能在1毫秒（10^{-2} 秒）内检查这252种二划分方式。如果我们接下来希望将一个有100个节点的网络划分成大小为$N_1=N_2=50$的两部分，根据公式**8.3** ，我们需要检查大约10^{29} 种划分方式，在同一台计算机上运行将耗时大约10^{16} 年。即使是在一个中等标度的网络中，检查所有可能的二划分方式都是不可行的，因此我们的穷举法一定会失败。

社区识别

在图划分问题中，社区的大小和数量是事先确定的。但在社区识别中，这二者都是未知的。若将网络分割成任意数目的组，使得每个节点属于且只属于一个组，我们将其称为划分。所有可能的划分的数量为[19-22]：

$$B_N = \frac{1}{e}\sum_{j=0}^{\infty}\frac{j^N}{j!} \qquad (8.6)$$

如图**8-7** 所示，当网络较大时，B_N 随网络节点数N 增长的速度高于指数级增长。

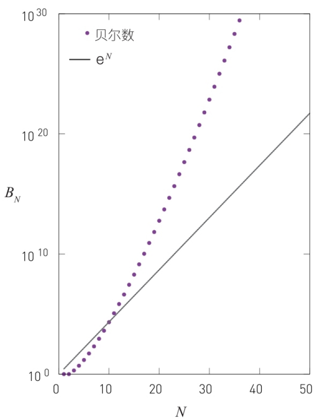


图**8-7**　划分的数量

有N 个节点的网络的划分的数量由贝尔数确定见公式**8.6** 。图中比较了贝尔数和指数函数，可能的划分数量的增长速度高于指数级增长。有$N=50$个节点的网络有10^{40}种划分，因此检查所有可能的划分方式来发现社区的穷举法在计算上是不可行的。

公式**8.5** 和**8.6** 指出了社区识别的根本挑战：把网络划分成社区的可能方式的数量随着网络规模N 呈指数级增长或更快。因此不可能检查一个大型网络的所有划分（边栏**8.2** ）。

边栏**8.2**

NP-完全性问题

运行一个算法需要多长时间？实际上，我们无法回答这个问题，因为运行时间取决于运行算法的计算机的速度。我们只能回答运行算法需要多少次计算。例如，寻找N 个数中最大数的算法需要将每个数与目前找到的最大数逐一比较，因此运行时间正比于N 。一般来说，如果一个算法的运行时间为N^x ，我们称之为多项式算法。

运行时间正比于N^3 的算法在任何计算机上都慢于运行时间为N 的算法。但与运行时间为2^N 的指数级算法相比，这个差异是很小的。例如，如果运行时间正比于N 的算法在$N=100$个元素上耗时1秒，那么N^3 算法在同一台计算机上耗时约3小时。然而一个指数级算法（2^N ）需要10^{20} 年才能算完。

能在多项式时间内通过算法解决的问题被称为P问题（class P problem）。虽然在网络科学中遇到的一些计算问题目前还没有多项式时间算法，只能用指数级运行时间的算法求解，但我们可以迅速在多项式时间内检查一个解是否正确。这类问题被称为NP-完全性问题，包括旅行商问题（图**8-8** ）、图着色问题、最大团发现、将图划分为指定类型的子图、点覆盖问题（边栏**6.4** ）。

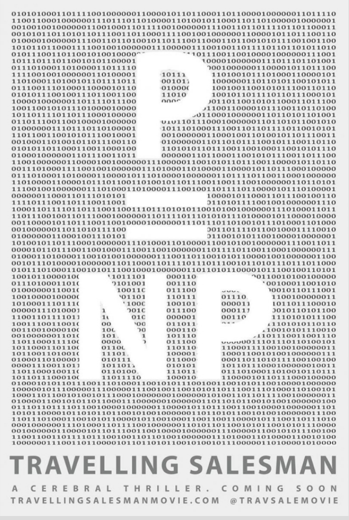


图8-8　电影之夜

《旅行商》（*Travelling Salesman*）是2012年的一部关于解谜的惊悚电影，讲述了4个数学家在解决了“P与NP问题”之后碰到的种种诡异问题。“P与NP问题”指“每一个能在多项式时间内求解的问题是否都能在多项式时间内求解”。这是7个千禧年难题之一，第一个正确回答的人可以获得100万美元奖金。“旅行商问题”是指一个商人想要以最短路径造访多个城市，每个城市只造访一次，最后回到出发的城市。虽然这个问题看起来很简单，实际上它是NP-完全性问题，寻找最短路径需要尝试所有组合。

NP-完全性问题在大众媒体中具有很高的有吸引力。美国哥伦比亚广播公司的系列电视剧《数字追凶》（*Numbers*）中的主角查理·艾普斯（Charlie Epps）因为相信某个NP-完全性问题的解可以治疗母亲的疾病，在母亲生命的最后3个月里一直试图求解这一问题。与之类似，该公司的另一系列电视剧《基本演绎法》（*Elementary*）中一次双重凶杀的动机也是为了寻找一个对密码学具有重大价值的NP-完全性问题的解。

总而言之，我们的社区概念来源于这样一种预期：每个社区都对应一个局部紧密连接的子图。这一假设给团、弱社区、强社区等多种不同的社区定义方式留下了空间。只要我们采用了一种定义，就可以通过检查网络中所有可能的划分并选择最适合这种定义的划分，来识别出社区。然而，划分的数量随网络节点数的增长呈指数级增长，甚至更快，使得穷举法在计算上不可行。因此，我们需要不必检查所有划分就能识别出社区的算法。这是接下来我们要讨论的主题。

8.3 层次聚类

若想识别大型真实网络的社区结构，我们需要运行时间为N多项式的算法。本节的主题层次聚类（hierarchical clustering）能帮助我们达到这一目标。

层次聚类首先计算相似度矩阵（similarty matrix），矩阵的每个

元素x_{ij}表示节点i和节点j之间的距离。在社区识别中，相似度刻画了节点i和j在网络中的相对位置。

当我们计算好x_{ij}之后，层次聚类迭代地将高相似度的节点识别为节点组。我们有两种不同的程序：凝聚算法（agglomerative algorithms）将高相似度的节点合并到同一社区，而分裂算法（divisive algorithms）会删除社区之间的低相似度链接以将社区彼此分离。两种程序都会生成一棵表示可能的社区划分的层次树，我们称之为树状图。接下来我们讨论凝聚算法和分裂算法的使用，以识别网络中的社区。

凝聚算法：劳沃斯算法

我们接下来以劳沃斯算法为例，介绍利用凝聚层次中的聚类算法进行社区识别。该算法用于识别代谢网络中的功能模块[11]，包括4个步骤。

步骤1：定义相似度矩阵

在凝聚算法中，属于同一社区的节点之间的相似度应该较高，属于不同社区的节点之间的相似度应该较低。在网络中，相互连接且有共同邻居的节点可能属于同一社区，因此x_{ij}应该较大。以下是拓扑重叠矩阵的公式（图**8-9**）[11]：

$$x_{ij}^{0} = \frac{J(i,j)}{\min\left(k_i, k_j\right) + 1 - \Theta\left(A_{ij}\right)} \tag{8.7}$$

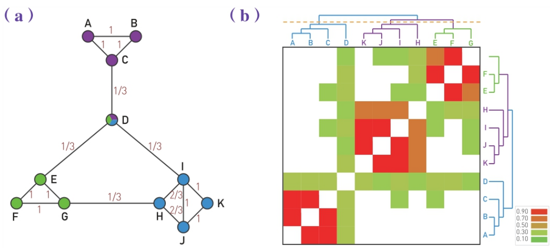


图**8-9**　劳沃斯算法

劳沃斯提出的凝聚分层聚类算法用于识别代谢网络中的功能模块，也可以用于任意网络。

（**a**）拓扑重叠

此图展示如何计算拓扑重叠x_{ij}^0 的小网络。对于每对节点i 和j ，根据公式**8.7** 计算其拓扑重叠。所有相邻节点的x_{ij}^0 标示在它们之间的链接上。注意，不相邻的节点的x_{ij}^0 可以是非零的。例如C和E的x_{ij}^0 =1/3。

（**b**）拓扑重叠矩阵

此图描述的是a图中所示网络的拓扑重叠矩阵x_{ij}^0 。矩阵的行和列已经在平均连接聚类后重新排列，使每个节点与拓扑重叠最高的节点相邻。颜色表示a图中计算的每对节点的拓扑重叠度。沿着橙色线分割这一树状图，可以看到网络中有三个模块。根据树状图所示，相比于ABC模块，EFG和HIJK模块之间更接近。[11]

刻画了这一预期（x_{ij}^0 上标0表示第0轮迭代）。式中Θ (x)是赫维赛德阶跃函数（Heaviside step function），当$x \le 0$时取值为0，当$x > 0$时取值为1；$J(i, j)$是节点i 和j 的公共邻居数，且当i 和j 之间直接相连时取值加1；$\min(k_i, k_j)$是度k_i 和k_j 中的较小者。因此：

- $x_{ij}^0=1$，若节点i 和j 直接相连且所有邻居都是公共邻居，如图**8-9a** 中的A和B。
- $x_{ij}^0=0$，若节点i 和j 没有公共邻居且并不相连，例如A和E。
- 处于局部连接稠密的网络邻域内的节点之间具有较高的拓扑重叠，例如节点H、I、J、K或节点E、F、G。

步骤**2**：确定节点组的相似度

当节点并入小的社区后，我们需要度量两个社区之间的相似度。有三种根据节点相似度矩阵x_{ij} 计算社区相似度的常用方法，分别是单一簇相似度（single cluster similarity）、完全簇相似度（complete cluster similarity）和平均簇相似度（average cluster similarity）（图**8-10** ）。劳沃斯算法使用平均簇相似度定义两个社区之间的相似度，其为分处两个社区的所有节点对i 和j 的x_{ij} 的平均值（图**8-10d** ）。

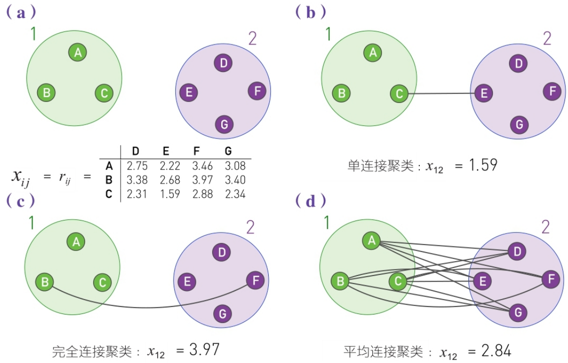


图8-10　簇相似度

在凝聚聚类中，需要根据节点相似度矩阵x_{ij} 确定两个社区之间的相似度。我们将这一过程形象地展示为空间中的一组点，点与点之间的物理距离r_{ij} 即为相似度x_{ij} 。在网络中，x_{ij} 表示某种基于网络的距离度量，例如公式**8.7** 中定义的x_{ij}^{0} $=0$。

（**a**）相似度矩阵

7个节点构成2个社区。表中给出了每对节点之间的距离r_{ij} ，即相似度x_{ij} 。

（**b**）单连接聚类

社区1和2之间的相似度定义为所有x_{ij} 中的最小值，其中i 和j 分属不同社区。因此相似度为$x_{12}=1.59$，即节点C和E之间的距离。

（**c**）完全连接聚类

两个社区之间的相似度定义为所有x_{ij} 中的最大值，其中i 和j 分属不同社区。因此$x_{12}=3.97$。

（**d**）平均连接聚类

两个社区之间的相似度定义为所有x_{ij} 中的平均值，其中i 和j 分属不同社区。这是劳沃斯算法采用的方法，$x_{12}=2.84$。

步骤3：执行层次聚类

劳沃斯算法使用以下步骤识别社区：

（**1**） 将每个节点视为一个社区，计算所有节点对之间的x_{ij} 。

（**2**） 寻找相似度最高的两个节点或两个社区，然后合并为一个社区。

（**3**） 计算新的社区与其他所有社区的相似度。

（4） 重复（1）和（2）直到所有节点都并入同一社区。

步骤4：树状图

步骤3的成对合并最终将所有节点并入同一社区。我们可以用树状图来选取合理的社区结构。

树状图展示了节点被归入具体社区的顺序。例如，图**8-9b** 的树状图告诉我们算法首先合并节点A和节点B，节点K和节点J，节点E和节点F，因为这些节点对的相似度为x_{ij}^{0} =1。接下来把节点C加入(A，B)社区，把节点I加入(K，J)社区，把节点G加入(E，F)社区。

若想识别社区，就需要将树状图分割。层次聚类并未告诉我们应该在哪里分割。例如，可以用图**8-9b** 中的短虚线来分割，我们会得到三个明确的社区（ABC、EFG和HIJK）。

劳沃斯算法在大肠杆菌代谢网络（图**8-3a** ）上识别出了细菌代谢的嵌套社区结构。为了检验这些社区的生物相关性，我们根据已知的每种代谢产物的生化分类，为树状图的各个分支标上了颜色。如图**8-3b** 所示，生化角色相近的基质倾向于出现在树上的同一分支中。换言之，已知的代谢产物的生化分类证实了网络拓扑中提取出的社区确实具有生物相关性。

计算复杂度

运行劳沃斯算法需要多少计算量？该算法有4个步骤，每一步的计算复杂度都不相同：

（**1**） 计算相似度矩阵x_{ij}^{0} 要求我们比较N^2 个节点对，因此计算量级为N^2。换言之，其计算复杂度为$O(N^2)$。

（**2**） 计算节点组的相似度要求我们在每一步要确定一个新簇与其他所有簇的距离。重复N 次需要$O(N^2)$次计算。

（**3**）（**4**） 构造树状图需要$O(N\log N)$步。

将步骤（1）～（4）合起来看，可知所需的计算步数量级为O (N

^{2})+O (N^2)+O ($N\log N$)。因为最慢的一步的量级为O (N^2)，算法的计算复杂度为O (N^2)。因此层次聚类远远快于穷举法，后者通常需要O (e^N)量级。

分裂算法：格文 - 纽曼算法

分裂算法系统地删除不同社区的节点之间的链接，最终将网络拆解成彼此独立的社区。我们通过米歇尔·格文和马克·纽曼提出的算法[9]，[23] 来介绍分裂算法的用途。该算法包括步骤。

步骤1：定义中心度

在凝聚算法中根据x_{ij} 选择节点进入同一社区，而在分裂算法中根据被称为中心度的x_{ij} 选择节点进入不同社区。因此我们希望x_{ij} 在节点i 和j 属于不同社区时较高，在节点i 和j 属于同一社区时较低。图**8-11**讨论了满足这一性质的三种中心度计算方式。其中计算最快的是链接介数(link betweenness)，定义x_{ij} 为通过链接(i ，j)的最短路径的数目。连接不同社区的链接的x_{ij} 应该较大，而社区内的链接的x_{ij} 应该较小。

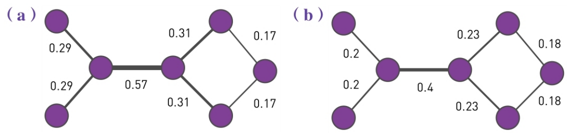


图**8-11**　中心度的计算方法

分裂算法要求属于不同社区的节点之间的中心度较高，属于同一社区的节点之间的中心度较低。两种常用的度量方式可以满足下面的一性质：

（**a**）链接介数

链接介数刻画了每条链接在信息传递中的作用。因此，x_{ij} 正比于所有节点对的最短路径中经过链接(i ，j)的最短路径数目。因此，社区之间的链接，例如图上正中的x_{ij} =0.57的链接，有较大的链接介数。计算链接介数的复杂度为O (LN)，在稀疏网络上为O (N^2) [23] 。

（**b**）随机游走介数

随机选择一对节点m 和n 。从节点m 开始，以均匀概率沿着每一条相邻链接游

走，直到抵达n 。随机游走介数x_{ij} 是游走经过链接$i \to j$ 的概率，遍历所有可能的m 和n ，取得平均值。计算随机游走介数要求对一个大小为$N \times N$ 的矩阵求逆，计算复杂度为$O(N^3)$，然后对所有节点对之间的游走取平均值，计算复杂度为$O(LN^2)$。因此随机游走介数的总计算复杂度为$O[(L+N)N^2]$，在稀疏网络上为$O(N^3)$。

步骤2：层次聚类

分裂算法的最终步骤与前述凝聚算法互为镜像（图8-12 ）：

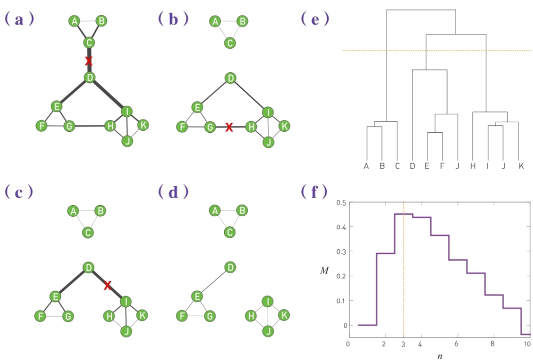


图8-12 格文 - 纽曼算法

（a） 格文和纽曼的分裂层次算法使用链接介数（图8-11a ）作为中心度。图中链接权重正比于x_{ij} ，可见连接不同社区的链接具有最高的x_{ij} 。事实上，社区之间的每一条最短路径都需要经过这些链接。

（b）～（d） 这一系列图形展示了算法如何逐一移除x_{ij} 最高的3条链接，使网络成为相互不连通的3个社区。注意每次移除链接之后需要重新计算介数。

（e） 该图是由格文 - 纽曼算法生成的树状图。在橙色虚线所示的第3级处分割可得网络中的3个社区。

（f）8 .4节 介绍的模块度函数M 可以帮助我们选择最优的分割。正如我们所期望的，其极大值给出的最优分割点在第3级，如图e所示。

（1） 计算每条链接的中心度x_{ij} 。

（2） 移除中心度最大的链接。如果多条链接有并列最大的中心度，随机选择其一。

（3） 重新计算剩下的网络中每条链接的中心度。

（4） 重复（2）和（3）直到所有链接都被移除。

格文和纽曼将这一算法应用于扎卡里空手道俱乐部网络（图**8-2a**），所得到的社区与原网络分解后的两组节点几乎完全一致，只有节点3被分错了。

计算复杂度

决定分裂算法速度的是计算中心度的步骤。因此，该算法的计算复杂度取决于我们选择哪种中心度。最快速的是链接介数，计算复杂度为$O(LN)$[24-26]（图**8-11a**）。

算法的步骤3在运行时间里多了一个因子L。因此算法的量级为$O(L^2N)$，在稀疏网络上为$O(N^3)$。

真实网络的层次结构

层次聚类提出了两个基础问题。

嵌套社区

首先，它假设小模块是嵌入大模块中的。树状图清楚地展示了这样的嵌套社区（图**8-9b** 和图**8-12e**）。然而，如何知道这样的层次结构真的存在于网络中呢？是否有这种可能，无论网络有没有嵌套社区结构，我们的算法都强制生成了这样的层次结构？

社区和无标度性质

其次，密度假设认为，一个网络可以被划分为一组相互弱链接的子图。若一个无标度网络存在枢纽节点，其必然连接多个社区，那么为什么会存在互不连通的社区呢？

图**8-13** 所示的层次网络模型解决了社区和无标度性质之间的冲突，并给出了嵌套层次社区结构的直观认识。这样得到的网络有几个关键特征。

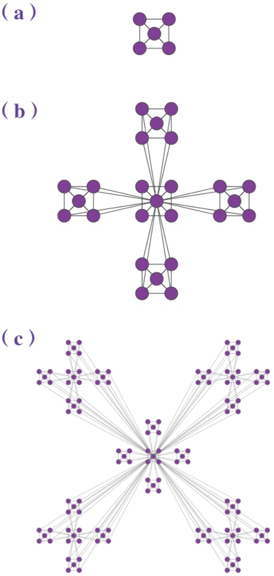


图**8-13**　层次网络

这三幅图展示了一个确定性的层次网络的迭代构造过程。

（**a**）此图从5个节点构成的全连通模块开始。注意对角线上的节点也是相连的，只不过链接没有画出。

（**b**）此图描述的是，构造与初始模块完全相同的4个模块，并将每个模块的外围节点连接到初始模块的中心节点。这样我们构造了一个有$N=25$个节点的网络。

（**c**）构造与这个25节点模块完全相同的4个模块，仍然将每个外围节点连接到初始模块的中心节点，得到有$N=125$个节点的网络。这一过程可以无限继续。[27]

无标度性质

层次模型生成具有以下度指数的无标度网络（图**8-14a**，进阶阅读**8.A**）：

$$\gamma = 1 + \frac{\ln 5}{\ln 4} = 2.161$$

与网络大小无关的集聚系数

虽然在埃尔德什－雷尼模型和巴拉巴西－阿尔伯特模型中集聚系数随N递减（5.9节），在层次网络中我们有$C=0.743$与网络大小无关（图8-14c）。我们还在代谢网络中观察到了这种与N无关的集聚系数[11]。

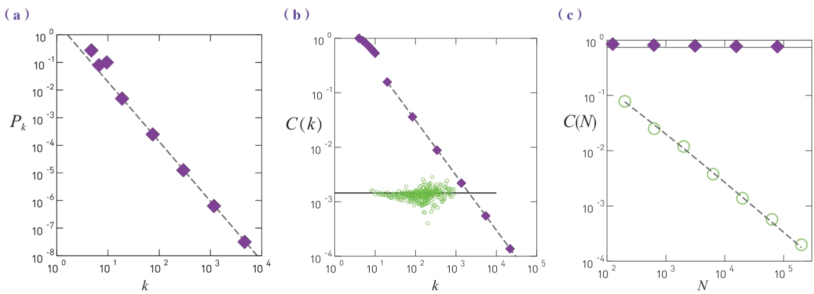


图8-14　层次网络的标度

图8-14表示了层次网络的三个特征量。

（a）度分布

p_k 的标度展示了生成的网络的无标度性质。斜率$\gamma=\ln5/\ln4$，如短虚线所示。进阶阅读8.A中有度指数的推导过程。

（b）层次聚类

$C(k)$服从公式8.8，如短虚线所示。圆圈表示对原网络进行标度随机化后的随机连接的无标度网络的$C(k)$。标度不存在，表明重连后丢失了层次结构。因此，$C(k)$表达了度分布以外的性质。

（c）与网络大小无关的集聚系数

集聚系数C与网络大小N的依赖关系。层次模型的C独立于N（实心记号），而巴拉巴西－阿尔伯特模型的$C(N)$则是递减趋势（空心记号）。[27]

层次模块度

该模型由很多小社区组成，小社区再组成较大的社区，进而组成更大的社区。这一嵌套层次模块的度的量化标志是节点的集聚系数依赖于节点的度[11]，[27]，[28]：

$$C(k)\sim k^{-1} \tag{8.8}$$

换言之，节点的度越高，其集聚系数越小。

公式**8.8** 刻画了社区在网络中的组织形式。事实上，小度节点的C 较大是因为它们处于密集的社区中。高度节点的C 较小是因为它们连接了不同的社区。例如，在图**8-13c** 中，在5个节点构成的模块的中心节点有$k=4$，且集聚系数$C=4$。在25个节点构成的模块的中心节点有$k=20$，且$C=3/19$。在125个节点构成的模块的中心节点有$k=84$，而且$C=3/83$。因此，节点的度越高，它的C 越小。

根据层次网络模型的这一性质，我们可以通过检查$C(k)$来判断一个网络是否具有层次结构。对于埃尔德什 - 雷尼模型和巴拉巴西 - 阿尔伯特模型，$C(k)$独立于k，不存在层次模块度。接下来，为了观察真实系统中是否存在层次模块度，我们在10个参照网络上计算$C(k)$，有以下几项发现（图**8-36**）：

- 只有电网不具有层次模块度，它的$C(k)$独立于k（图**8-36a**）。
- 余下的9个网络的$C(k)$随k 递减。因此在这些网络中，小节点属于小的密集社区，而枢纽节点将不同的社区连接起来。
- 在科学合作网络、代谢网络和引文网络中，$C(k)$在k 较大时服从公式**8.8**。互联网、手机网络、电子邮件网络、蛋白质相互作用网络、万维网的$C(k)$不服从公式**8.8**，需要分别推导。更精细的网络模型导出$C(k)\sim k^{-\beta}$，其中β 介于0和2之间[27]，[28]。

总而言之，从原则上来说，层次聚类不需要事先知道社区的数目和大小。实践中层次聚类生成树状图，以给出目标网络的社区划分谱系。树状图并未告诉我们如何划分能最好地刻画潜在社区结构。事实上，对层次树的任意分割都给出了一种可能的有效划分（图**8-15**）。这与我们的期望并不吻合——我们通常认为一定存在标准答案，对应唯一的社区结构。

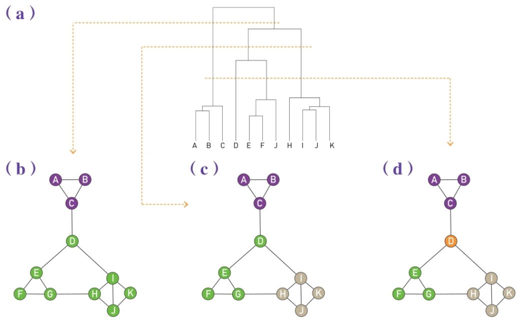


图8-15　层次聚类的模糊性

层次聚类并没有告诉我们应该在哪里分割树状图。事实上，在不同位置分割图**8-9a**的树状图，我们会得到2个（图b）、3个（图c）或4（图d）个社区。虽然对于小网络，我们可以读图判断哪种分割最优地刻画了潜在的社区结构，但对于大型网络却完全做不到。下一节中讨论的模块度可以帮助我们选择最优分割。

网络中有不少关于层次结构的概念[29]，[30]，而检查$C(k)$可以帮助我们判断一个网络是否有层次模块度。我们发现在大多数真实网络中$C(k)$是递减趋势，这表明大多数真实系统中存在层次性模块度。同时，在埃尔德什-雷尼模型和巴拉巴西-阿尔伯特模型中，$C(k)$独立于k，表明这些标准模型中不具有层次组织。

8.4 模块度

在随机连接的网络中，节点之间的连接模式应该是相同的，与网络的度分布无关。因此，这些网络不应该表现出系统性的局部密度差异，也就是我们所称的社区。这催生了社区组织的第三个假。

H3：随机假设

随机连接的网络没有内在社区结构。

该假设对实践的指导意义在于，通过比较社区的链接密度和随机网络（对这个社区内的节点进行随机重连）的链接密度，可以判断出

这个社区是真实存在的密集子图，还是偶然生成的连通模式。

本节我们定义了物理量——模块度。它是指与随机连接方式的偏离，可以用于度量划分的优劣。所以模块度允许我们判断一种社区划分是否强于另一种。最后，优化模块度是一种新型的社区识别方法。

模块度

我们先来一个有N 个节点和L 条链接的网络。想象有一种划分，可以得该网络划分成n_c 个社区，每个社区有Nc 个节点，节点之间用L_c 条链接相连，其中$c = 1,\ldots,n_c$ 。给定网络的度序列。根据H2密度假设的推论，若L_c 大于N_c 个节点之间链接的期望数，那么子图C_c 的节点可能确实是一个真实社区的一部分（图**8-2** ）。因此，我们可以度量网络的真实连接模式（A_{ij} ）和随机连接时i 和j 的期望链接数（p_{ij} ）之间的差异：

$$M_c = \frac{1}{2L}\sum_{(i,j)\in C_c}(A_{ij} - p_{ij}) \quad (8.9)$$

公式中的p_{ij} 可以通过随机化原网络来计算，同时保持每个节点的度的期望不变。根据保度零模型（degree preserve null model）（公式**7.1** ），我们可以得出：

$$p_{ij} = \frac{k_i k_j}{2L} \quad (8.10)$$

若M_c 为正，那么子图C_c 的链接多于随机连接的期望链接数，则它可能是一个社区。若M_c 为零，那么这N_c 个节点之间的连接是随机的，可以由度分布完全解释。最后，若M_c 为负，那么C_c 的节点不构成社区。

基于公式**8.10** 我们可以推导出模块度（公式**8.9** ）的简化形式（进阶阅读**8.B** ）

$$M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 \quad (8.11)$$

其中L_c是社区C_c内的链接数，k_c是该社区的节点的度的总和。

接下来我们将这样的想法扩展到整个网络。考虑一个完整的划分（complete partition），把网络划分为n_c个社区。我们定义这个划分的模块度为公式**8.11** 在所有n_c 个社区上进行的求和，以观察该划分定义的各个子图的局部链接密度与随机连接网络的期望密度之间的差异[23]：

$$M = \sum_{c=1}^{n_c}\left[\frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2\right] \tag{8.12}$$

模块度有几个关键性质。

- 高模块度表示好的划分

划分的M 越高，对应的社区结构就越好。事实上，图**8-16a** 中的划分的模块度最大（$M = 0.41$），准确地划分出两个显而易见的社区。模块度较低的另一种划分则明显与真实社区有所差异（图**8-16b**）。注意一个划分的模块度不能超过1[31]，[32]。

- 模块度为零或者为负的情况

将整个网络视为单一社区可得$M = 0$，这时公式**8.12** 括号里的两项相等（图**8-16c** ）。如果将每个节点都视为单独的社区，我们将得出$Lc = 0$，且公式**8.12** 中的n_c 项都为负，因此M 为负（图**8-16d**）。

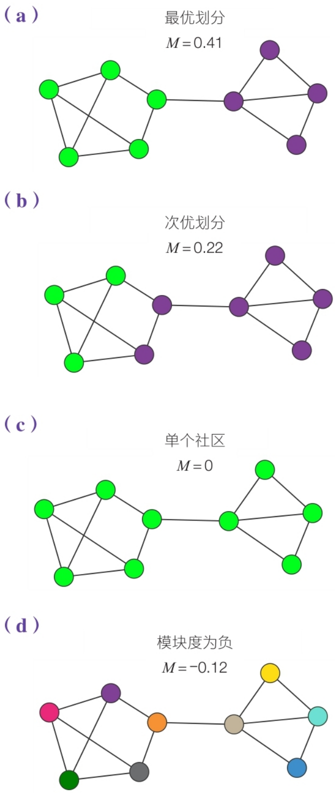


图8-16　模块度

图中的网络有两个显而易见的社区。我们根据公式**8.12**　的定义给出几种划分的M值，以便直观展示模块度的意义。

（**a**）最优划分

模块度最大的划分，$M=0.41$与两个社区吻合。

（**b**）次优划分

模块度不是最大但为正的划分，$M=0.22$，未能正确识别网络中的社区

（**c**）单个社区

在任何网络结构中，令所有节点属于同一个社区，可得$M=0$。

（**d**）模块度为负

如果我们把每个节点单独分配到不同的社区，模块度为负，可得$M=-0.12$。

我们可以用模块度来判断，层次聚类方法给出的若干个划分中哪

个是社区结构最优的，答案是M 最大的那个。图**8-12f** 展示了树状图每个分割对应的M 值，显然把网络分成三个社区的分割方式M 值最大。

贪婪算法

模块度越高的划分方式越能准确地识别潜在社区结构，这一认识让我们提出最后一条假设。

H4：最大模块度假设

给定一个网络，模块度最大的划分对应最优的社区结构。

在前述对小网络的考察中，模块度M 最大的划分确实与预设的社区一致。这些实验结果很好地支持了这一假设（图**8-12** 和图**8-16**）。

最大模块度假设为一些社区识别算法提供了基础。这些算法试图寻找模块度最大的划分。原则上我们可以对所有可能的划分计算M ，然后找出M 最大的划分作为最优划分。然而，由于划分的数量非常大，穷举法在计算上不可行。接下来我们讨论一种算法，这种算法不必检查所有的划分就可以找到接近最大M 的划分。

贪婪算法

第一个最大化模块度算法由马克·纽曼提出[33]，这种算法迭代地将社区两两合并，直至无法增大模块度。算法的步骤如下：

（**1**） 将每个节点划为一个单独的社区，即一开始有N 个单点社区。

（**2**） 对于有至少一条链接相连的两个社区，计算合并两个社区前后模块度的改变ΔM 。检查所有这样的两个社区，找到ΔM 最大的一对，合并它们。注意模块度始终是在整个网络上计算的。

（**3**） 重复步骤（2）直到所有节点合并入同一个社区，并记录每一步的M 。

（4）选择M 最大的划分。

为了展示贪婪算法的识别能力，请想象一个由物理学家组成的合作网络，包括在arxiv.org上贴出论文的所有N ＝56 276名科学家，覆盖了所有物理领域（图**8-17** ）。贪婪算法在M 最大时（M ＝0.713）识别出大约600个社区。其中4个社区非常大，合起来囊括了全部节点的77%（图**8-17a** ）。最大社区的93%的作者在凝聚态物理领域发表论文，第二大社区的87%的作者在高能物理领域发表论文。这表明每个社区内的物理学家的专业兴趣彼此相近。图**8-2a** 也展示了贪婪算法的准确性，扎卡里空手道俱乐部网络的最大M 的社区结构与该俱乐部随后的分裂相吻合。

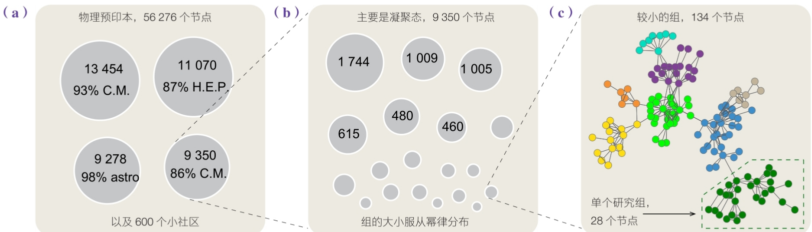


图**8-17**　贪婪算法

（**a**）对物理学家聚类

这三幅图描述的是物理学家合作网络的社区结构。贪婪算法识别出4个大社区，每个社区都主要由有相似兴趣的物理学家构成。我们在每个簇上标示了属于同一物理领域的成员比例，并根据物理学家在线贴出预印论文时所选择的分类来标记他们的专业兴趣。C.M.代表凝聚态物理，H.E.P.代表高能物理，astro代表天体物理。这4个大社区和600个小社区共同构成了总模块度为M ＝0.713的划分。

（**b**）识别子社区

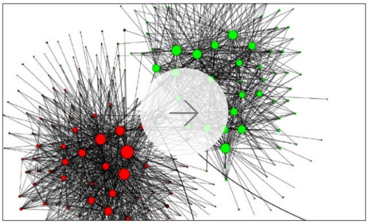
我们可以在每个社区内使用贪婪算法，将社区视为单独的网络，从而识别子社区。这一过程把凝聚态社区分割成了较小的子社区，将划分的模块度提高到了M ＝0.807。

（**c**）研究组

对其中一个小社区进一步划分，可以看到一个个物理学家和他们所属的研究组。[33]

计算复杂度

因为计算每个ΔM 可以在常数时间完成，贪婪算法的步骤（2）需要$O(L)$次计算。在决定合并哪两个社区后，更新矩阵所需时间在最坏的情况下为$O(N)$。因为算法需要$N-1$次社区合并，其复杂度为$O[(L+N)N]$，在稀疏图上为$O(N^2)$。对算法进行优化后可以将复杂度降低到$O(N\log^2 N)$（在线资源**8.1**）。



在线资源**8.1**　基于模块度的算法

一些广泛使用的社区识别算法的目标是最大化模块度。

优化的贪婪算法

使用稀疏矩阵的数据结构可以将贪婪算法的计算复杂度降低到$O(N\log^2 N)$[35]。

Louvain算法

模块度优化算法的计算复杂度降到了$O(L)$[2]。因此我们得以在上百万节点上识别社区，如在线资源**8.1** 所示。该算法详见进阶阅读**8.C**。

扫描二维码下载“湛庐阅读”App，回复书名，详细了解模块度的算法。

模块度的局限

鉴于模块度在社区识别中的重要作用，我们必须意识到它存在的局限。

分辨率极限

最大化模块度使得小社区必须被合并入大社区[34]。事实上，如

果我们把社区A和社区B合并为一个社区，网络的模块度会变为（进阶阅读**8.B**）：

$$\Delta M_{AB} = \frac{l_{AB}}{L} - \frac{k_A k_B}{2L^2} \quad (8.13)$$

其中l_{AB}是社区A中节点和社区B中节点之间的链接数，社区A中所有节点的度之和为k_A，社区B所有节点的度之和为k_B。若A和B是完全不相连的社区，那么最大化M后，它们依旧分属不同社区。如我们接下来所示，A和B完全不相连的情况并不总会发生。

考虑$\frac{k_A k_B}{2L} < 1$的情况。根据公式**8.13**，若两个社区之间存在至少一条链接（$l_{AB} \geq 1$），则有$\Delta M_{AB} > 0$。因此，要最大化模块度就必须合并A和B。简化考虑，假设$k_A \sim k_B = k$，若这两个社区的节点的度之和满足以下公式：

$$k \leqslant \sqrt{2L} \quad (8.14)$$

那么即使A和B完全不相连，合并A和B依然会导致模块度增加。这是模块度最大化的产物：若k_A和k_B确实低于公式**8.14**的阈值，这两个社区之间的链接的期望值小于1。因此，即使两个社区之间只存在一条链接，也会在最大化M的时候导致两个社区合并。分辨率极限导致了如下后果：

- 模块度最大化的方法无法检测到小于分辨率极限（公式**8.14**）的社区。例如，在万维网样本（表**1-1**）中L = 1 497 134，模块度最大化很难观察到节点度之和$k_C \lesssim$ 1 730的社区。
- 真实网络中有大量小社区[36-38]。根据分辨率极限（公式**8.14**），这些小社区会被系统地并入大社区，导致我们对社区结构的认识有所偏差。

若想要避免分辨率极限的问题，我们可以进一步划分优化模块度得到的大社区[33]，[34]，[39]。例如，将图**8-17a**里的两个凝聚态社区中的较小者视为独立网络，对它再执行一遍贪婪算法，可以得到100

个小社区，模块度亦增大到$M=0.807$（图**8-17b**）[33]。

模块度的最大值

所有基于最大化模块度的算法都依赖一个假设：有清晰社区结构的网络一定存在对应最大M的最佳划分[40]。在实际操作中，我们希望最大值M_{max}易于寻找，并且它给出的社区应迥异于所有其他划分给出的社区。然而，正如下面的内容所示，我们很难从一大群接近最优的划分中找到最优划分。

考虑一个网络有n_c个子图，链接密度$k_c \approx 2L/n_c$。最优划分应该把每个子图划为一个单独社区（图**8-18a**），对应的模块度为$M=0.867$。然而，如果把相邻子图两两合并为一个社区，我们会得到一个更高的模块度$M=0.87$（图**8-18b**）。一般来说，根据公式**8.13**和**8.14**，如果我们合并两个子图，对模块度的改变为：

$$\Delta M = \frac{l_{AB}}{L} - \frac{2}{n_c^2} \qquad (8.15)$$

换言之，模块度的变化量小于$\Delta M=-2/n_c^2$。对于有$n_c=20$个社区的网络来说，模块度的变化量最多为$\Delta M=-0.005$，与最大模块度$M=0.87$（图**8-18b**）相比非常微小。随着社区数量增加，ΔM_{ij}趋近于零，因此找到最优划分会越来越难，因为最优划分和大量的次优划分的模块度几乎一样。换言之，模块度函数并不是在单一最优划分处存在峰值，而是存在一个由高模块度值构成的高台（图**8-18d**）。

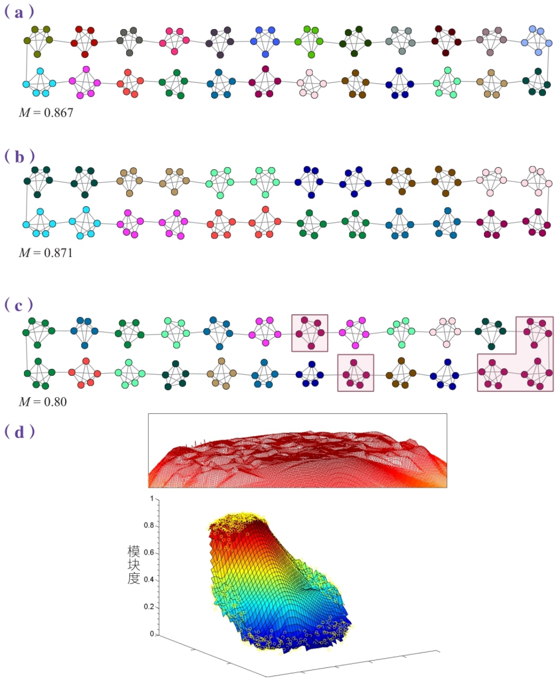


图8-18　模块度的最大值

有24个团的环形网络，每个团有5个节点。

（**a**）直观划分

最好的划分应该与网络配置一致，每个团簇是一个单独社区。这一划分的模块度为$M=0.867$。

（**b**）最优划分

如果如图中节点颜色所示，把团簇两两合并，我们会得到$M=0.871$，高于直观划分所得的模块度。

（**c**）随机划分

模块度相近的划分可以有截然不同的社区结构。如果我们对每个团簇随机地分配社区标签，即使不相连的团也允许划入同一社区，例如图中5个高亮的团簇被划入同一社区。这一随机划分也有很高的模块度$M=0.80$，与最优值$M=0.87$相差无几。

（**d**）模块度高原

根据997种划分方式给出的网络（图a）的模块度函数值。纵轴表示模块度M，可以看到一个具有高模块度的高原，周围是大量的低模块度的划分。因此，并没有一个明确的模块度最大值。此时，模块度函数是高度退化的。[40]

总而言之，模块度给出了理解网络社区结构的第一原理。事实上，它把一系列基本问题表达成一个简洁的形式，包括如何定义社区，如何选择合适的零模型，以及如何度量一个划分的优劣。因此，最优化模块度在社区识别领域中扮演着核心角色。

同时，模块度有一些众所周知的局限。首先，它将弱链接的小社区强行合并。其次，网络通常没有明确的模块度最大值，而是存在一个模块度高台。高台区中各个划分的模块度都很高，难以区分。高台区解释了为什么很多最大化模块度算法能快速找到一个高M值的划分：它们其实是找到了接近最优M的诸多划分中的一个。最后，解析计算和数值模拟表明，即使是随机网络也存在高模块度的划分，这与随机假设H3矛盾，而模块度的概念正是源自后者[41-43]。

模块度优化是一个更大问题的特例：通过优化某种品质函数（quality function）Q 来寻找社区。贪婪算法和进阶阅读**8.C** 中介绍的Louvain算法假设$Q=M$，并寻找模块度最大的划分。进阶阅读**8.C**中也会介绍Infomap算法，通过最小化映射方程L 来寻找社区。L 是基于熵的划分品质度量[44-46]。

8.5 重叠社区

很少有节点会只属于单一社区。比如一位科学家，他属于有共同科研兴趣的科学家社区，也属于由家人和亲戚构成的社区，可能还属于有共同业余爱好的人的社区（图**8-19**）。这些社区的每个成员都属于多个其他社区，构成了有嵌套和重叠社区的复杂网络[36]。重叠社区并不限于社交网络：同一种基因会涉及多种疾病，表明不同症状的疾病模块之间有重叠[14]。

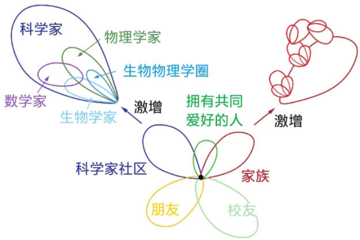


图8-19　重叠社区

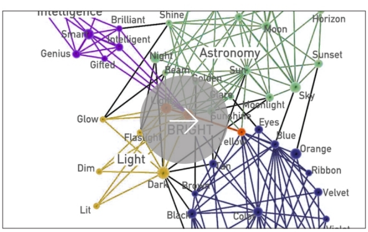
此图是引入重叠社区概念的陶马什·维切克本人的社区的示意图。对科学家社区进行放大观察，我们可以看到表征他科研兴趣的社区的嵌套和重叠结构。[36]

尽管社会学家和对社区工程感兴趣的人对嵌套社区结构认识已久[47]，目前我们讨论的算法也要求每个节点属于单一社区。其转折点始于陶马什·维切克（Tamás Vicsek）和合作者们的工作[36]，[48]。他们提出了识别重叠社区的一种算法，让网络科学研究者们开始关注这个问题。本节我们会讨论两种发现重叠社区的算法：团渗流（clique percolation）和链接聚类（link clustering）。

团渗流

团渗流算法通常被称为CFinder，将社区视为重叠团的集合[36]：

- 两个k -团，如果有k -1个公共节点，我们则认为这两个团是相邻的（图**8-20b**）。
- k -团社区是由所有相邻k -团的集合构成最大连通子图（图**8-20c**）。



在线资源8.2　团渗流算法

团渗流算法让我们得以识别重叠社区。该软件可以从www.cfinder.org下载。

扫描二维码下载“湛庐阅读” App，回复书名，了解什么是团渗流算法。

- 从一个k -团出发无法到达的那些属于其他k -团社区的k -团（图**8-20c**、**d** ）。

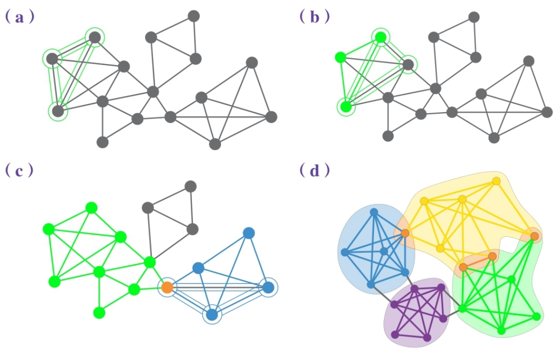


图8-20　团渗流算法

为了识别k =3的团构成的社区，我们让一个三角形在网络上滚动，使得每次滚动到下一个位置的三角形都与上一个位置的三角形共享一条链接（两个节点）。

（a） ~（**b**）滚动的团

从图a中所示的绿色三角形开始，算法下一步到达图b中所示的三角形。

（c） k =**3**的团构成的社区

此图描述的是绿色社区添加了最后一个三角形之后，算法的暂时结果。所有绿色三角形都不再能找到共有一条链接的新的三角形，绿色社区就此完成。注意，给定k 之后，同一网络中可以有多个通过团渗流得到的社区。第二个这样的社区如图中蓝色所示。图中高亮了蓝色社区添加了最后一个三角形之后的结果。蓝色和绿色社区有重叠，共有橙色节点。

（d） k =**4**的团构成的社区

一个示例网络中k =4的团构成的社区，每个社区都是由4个节点构成的完全子图组成，并且同一社区的各个子图都共有至少3个节点。橙色节点属于多个社区。感谢盖尔盖伊·帕拉（Gergely Palla）提供图片。

团渗流算法的步骤是，首先找出所有的团，然后建立$N_{\text{clique}} \times N_{\text{clique}}$团重叠矩阵$O$，其中$N_{\text{clique}}$是团的个数，$O_{ij}$是团$i$和团$j$共有的节点数（图**8-39**）。图**8-21** 给出了一个团渗流算法的典型结果：单词bright的社区结构。在这个网络中，两个单词如果有相关联的词义，二者之间就存在一条链接。我们很容易看到，该算法找到的重叠社区是有意义的，单词bright同时属于4个社区：一个与光线有关的单词社区，例如“发光”（glow）和“黑暗”（dark）；一个与颜色有关的单词社区，例如“黄色”（yellow）和“棕色”（brown）；一个与天文术语有关的单词社区，例如“太阳”（sun）和“射线”（ray）；一个与智慧有关的单词社区，例如“天才”（gifted）和“卓越”（brilliant）。这个例子也指出了此前的算法识别网络社区时遇到的困难：那些算法总会要求bright只属于4个社区的某一个，必须从另外3个中移除。因此，另外3个社区会丢失一个关键成员，导致社区的意义难以解释。

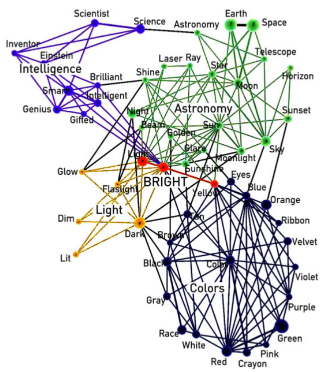


图8-21 重叠社区

南佛罗里达词联想网络中包含单词bright的社区。在这个网络中，节点是单词，词义相关联的单词之间有链接。团渗流算法识别出的社区结构描述了bright的多重词义，可以用于表示光线、颜色、天文术语或智慧的相关意思。[36]

有没有可能，团渗流识别的社区纯属偶然？为了区分k-团社区与仅仅由于高链接密度偶然产生的社区，我们将考察随机网络的k-团的渗流属性[48]。如我们在第**2**章讨论的，如果一个随机网络足够密集，它会有大量的不同大小的团。仅当连接概率p超过以下阈值时，

一个随机网络中才会出现大的k -团社区（进阶阅读**8.D** ）

$$p_c\langle k\rangle = \frac{1}{\left[(k-1)N\right]^{1/(k-1)}} \qquad (8.16)$$

当连接概率p 小于$p_c\langle k\rangle$ 时，我们认为只会出现少量孤立的k -团（图**8-22a** ）。一旦p 超过$p_c\langle k\rangle$ ，我们会观察到有大量的团组成了k -团社区（图**8-22b** ）。换言之，每个k -团社区有自己的阈值。

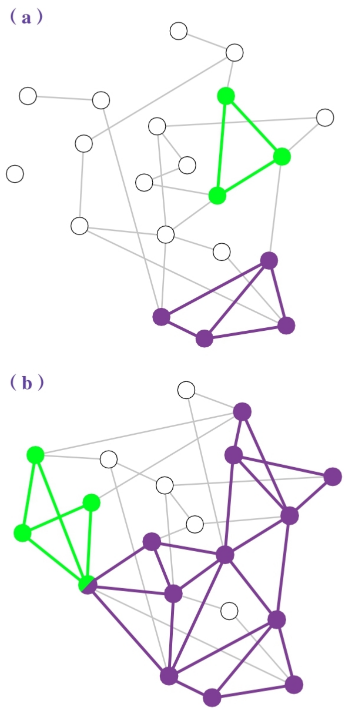


图**8-22**　团渗流算法

分别以概率p =0.13（图**8-22a** ）和p =0.22（图**8-22b** ）建立的随机网络。两个p 都大于链接渗流阈值（当N =20时，p_c =1/N =0.05），因此两个网络中大多数节点都属于巨连通分支。

（**a**）次临界社区

3-团（三角形）渗流阈值为p_c (3)=0.16，则p =0.13小于该阈值。因此只能观测到两个小的3-团渗流簇，且彼此不相连。

（**b**）超临界社区

p =0.22，高于p_c(3)，因此我们观测到多个3-团组成了一个很大的3-团渗流簇（紫色）。网络中还有第二个3-团社区与之重叠，以绿色表示。[48]

当k =2时，k -团相互连接，公式**8.16** 简化为$p_c\langle k\rangle \sim 1/N$ ，即为埃尔德什 - 雷尼网络产生巨连通分支的条件。

当k =3时，团即为三角形（图**8-22a**、**b**），公式**8.16** 给出$p_c\langle k\rangle \sim 1/\sqrt{2N}$ 。

换言之，足够密集的网络中会自然产生k -团社区。因此，要解释网络的重叠社区结构，我们必须将其与原网络度随机化后的社区结构相比较。

计算复杂度

寻找网络中的团的算法运行时间随N 指数增长。然而，团渗流算法的社区定义在于团，而非最大团，这可以在多项式时间内找到[49]。然而，如果网络中有很大的团，使用$O(e^N)$复杂度的算法来识别所有的团更为高效[36]。虽然计算复杂度很高，但是算法非常快，处理400万手机用户的手机通话网络只需要不到一天的时间[50]（图**8-28**）。

链接聚类

虽然节点通常属于多个社区，链接常常只属于单一社区，表达成员之间的确切关系，而这种关系定义了成员为什么属于一个社区。例如，两个人之间的链接要么代表他们在同一个家庭，要么代表他们在一起工作，或者表示他们有共同的业余爱好，很少会有重叠。与之类似，在生物学中，蛋白质的每种结合作用负责一种不同的功能，唯一地定义了蛋白质在细胞中的作用。链接的这一特性让社区识别算法发展出了对链接聚类而非对节点聚类的方法[51]，[52]。

由阿恩（Ahn）、贝格罗（Bagrow）和莱曼（Lehmann）提出的链接聚类算法[51] 包括两个步骤。

步骤1：定义链接相似度

两条链接的相似度由它们的端点的邻域确定。例如，考虑链接(i，k)和(j，k)共有同一个端点k，则它们的相似度定义为（图**8-23a**～**c**）：

$$S\left((i,k),(j,k)\right)=\frac{\left|n_{+}(i)\cap n_{+}(j)\right|}{\left|n_{+}(i)\cup n_{+}(j)\right|} \quad (8.17)$$

其中$n_{+}(i)$是节点i的邻居列表，包括其自身。则S度量了i和j的公共邻居的相对数。因此，若i和j有完全相同的邻居，则$S=1$（图**8-23c**）。两条链接的邻域的重叠越少，S就越小（图**8-23b**）。

步骤2：层次聚类

相似度矩阵S让我们得以使用层次聚类来识别链接社区（**8.3**节）。我们采用单链接方法，每轮合并相似度最大的两个链接社区（图**8-10**）。

综合起来，对于图**8-23e**所示的网络，公式**8.17**给出了图**8-23d**中所示的相似度矩阵。单链接层次聚类给出了图**8-23d**中所示的树状图，其分割给出了图**8-23e**中所示的链接社区和图**8-23f**中所示的重叠节点社区。

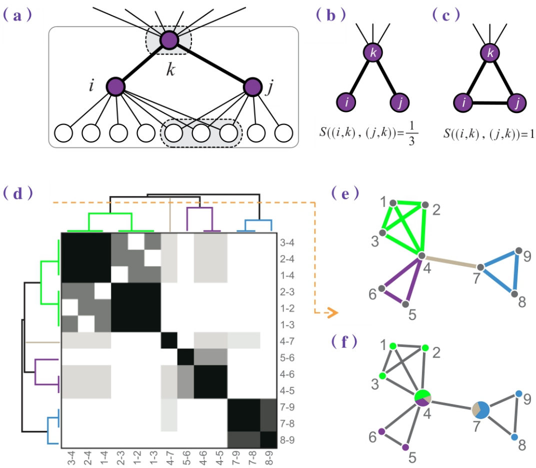


图8-23　识别链接社区

链接聚类算法用于分析每条边两个端点的连接模式，以找出网络中具有相似拓扑作用的链接。采用与劳沃斯算法[4]（图8-19）类似的函数，该算法的目标是对连接社区内节点的链接给予高相似度S。

（a） 共享端点k的链接(i, k)和(j, k)的相似度S检测这两条边是否属于同一社区。以$n_+(i)$表示节点i的邻居列表，包括其自身。我们有$|n_+(i)\cap n_+(j)|=12$且$|n_+(i)\cup n_+(j)|=4$，根据公式8.17可得$S=1/3$。

（b）对于存在孤立端的三个相连节点（$k_i=k_j=1$），我们有$S=1/3$。

（c）对于三角形，我们有$S=1$。

（d） 图e、f所示网络的链接相似度矩阵。颜色较深的元素对应相似度S较高的一对链接。图上也给出了最终的链接树状图。

（e）图d中树状图橙色虚线表示的分割给出的链接社区结构。

（f）图e中的链接社区结构导出的重叠节点社区。[51]

图8-24　展示了链接聚类算法在维克多·雨果的小说《悲惨世界》中的人物之间识别出的社区结构。任何熟悉这部小说的读者都可以确认，这些社区准确地表达了每个人物的身份。有几个人物置于多个社区中，反映了他们在小说中的多重身份。然而，每个社区的链接是这个社区独有的。

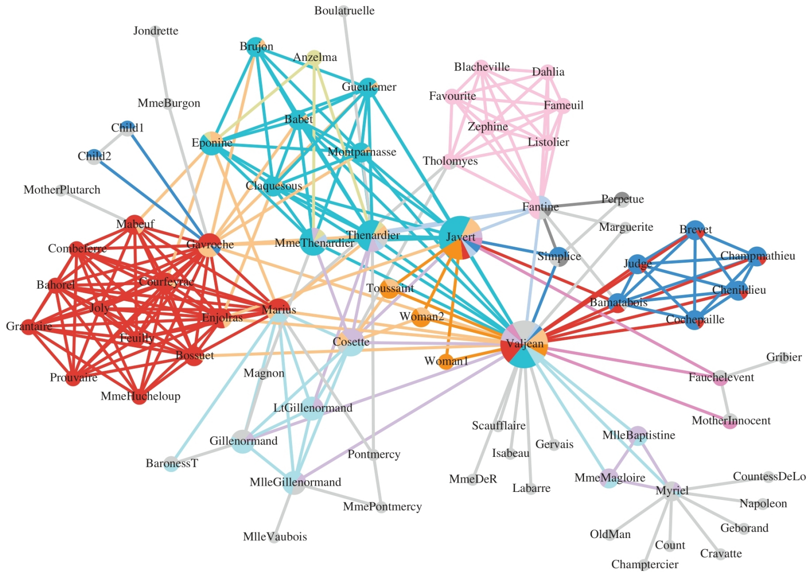


图**8-24**　链接社区

维克多·雨果小说《悲惨世界》（1862）中的人物网络。如果两个人物在故事中有直接交流，则他们的节点会相连。链接颜色表示簇，浅灰色节点对应单一链接的簇。属于多个社区的节点以饼图表示，展示人物在每个社群里的身份。毫无意外，主角冉·阿让的社区身份最多。[51]

计算复杂度

链接聚类算法包括两个耗时的步骤：相似度计算和层次聚类。对度为k_i和k_j的两条链接计算相似度（公式**8.17**）需要$\max(k_i，k_j)$步。对于度指数γ的无标度网络，计算相似度的复杂度为$O\left(N^{2/(\gamma-1)}\right)$，取决于最大节点的度$k_{max}$。层次聚类要求$O\left(L^2\right)$个时间步。因此，算法的总计算复杂度为$O\left(N^{2/(\gamma-1)}\right)+O\left(L^2\right)$。对于稀疏图，后一项起主要作用，则为$O(N^2)$。

识别重叠社区的需求催生了很多种算法[53]。例如，团渗流算法被扩展到分析加权[54]、有向和二部图[55]，[56]。与之类似，我们可以像**8.4**节讨论的模块度函数一样推导出链接聚类的评价函数[52]。

总而言之，本节讨论的算法承认节点应该天然属于多个社区。因

此，如果像前几节一样强制要求每个节点只能属于单一社区，我们得到的社区结构便是有错误的。链接社区认识到了这样一个事实：每条链接准确地反映了两个节点之间关系的本质。链接聚类也给出了网络中重叠社区的结构。

8.6 检验社区

社区识别算法提供了强大的分析工具，让我们得以描述真实网络的局部结构。然而，若想解释和应用找到的社区，还需要知道算法的准确率。与之类似，分析大型网络需要解决算法的计算效率。本节我们关注评估社区识别准确率和速度的概念。

准确率

如果网络连接图表达的社区结构是唯一的，那么每个算法都应该精确地找到同样的社区。然而，考虑到不同算法采用的不同假设，它们找到的划分可能是不同的。这提出了一个问题：我们应该采用哪种社区识别算法？

为了评估社区识别算法的性能，我们需要度量算法的准确率，也就是在已知社区结构的网络中找到社区的能力。我们以两个基准网络为例开始讨论。这两个网络有预先指定的社区结构，可以用于检验社区识别算法的准确率。

GN基准网络（Girvan-Newman）

GN基准网络包括$N=128$个节点，被划分为$n_c=4$个社区，每个社区大小为$N_c=32$ [9]，[57]。每个节点以概率p^{int}连接到社区内的N_c-1个节点，以概率p^{ext}连接到其他三个社区的$3N_c$个节点。控制参数为：

$$\mu = \frac{k^{\text{ext}}}{k^{\text{ext}} + k^{\text{int}}} \tag{8.18}$$

刻画了社区内外的密度差异。当μ 较小时，也就是连接到社区内部节点的概率高于连接到另一社区节点的概率时（图**8-25a**），我们希望社区识别算法有良好的性能。当μ 较大时，也就是社区内部的链接密度与网络中其他部分的链接密度相近时（图**8-25b**），所有算法将会失效。

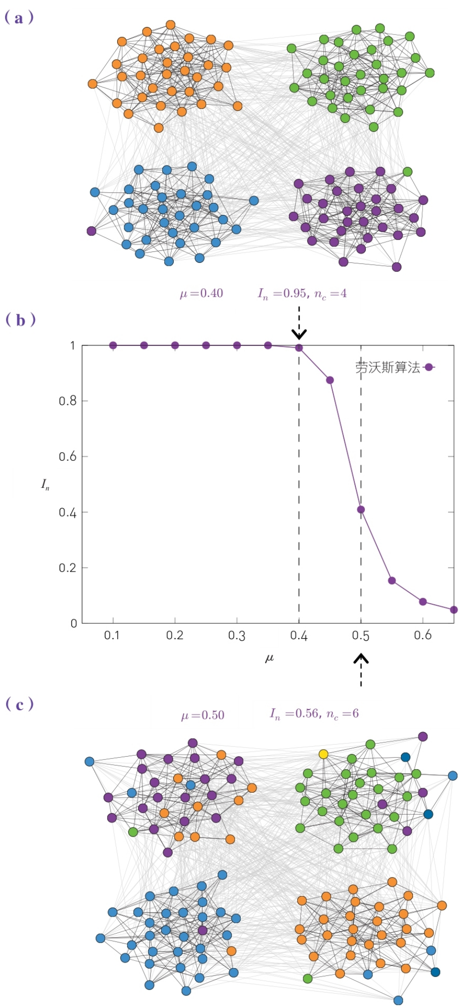


图8-25　使用GN基准网络检验准确率

图a和图c中通过节点的位置展示了GN基准网络的预设社区。4个不重叠的社区各自有N_c ＝32个节点。

（**a**） 节点颜色表示在公式**8.18** 所定义的混合参数μ ＝0.40时劳沃斯算法给出的划分。因为这种情况下的社区是良好分隔的，我们看到预设的社区与识别的社区吻合得很好。

（**b**） 归一化互信息关于劳沃斯算法混合参数μ 的函数曲线。当μ 较小时，我们有I_n ≈1且n_c ≈4，表明算法容易识别良好分隔的社区，如图a所示。当我们增大μ 时，社区内外的链接密度的差异变得不明显。因此社区难以发现，I_n 减小。

（**c**） 当μ ＝0.50时，劳沃斯算法错分了大量节点，因为这时社区不再是良好分隔的，很难发现正确的社区结构。

我们注意到劳沃斯算法生成了多个划分，因此对于每个μ ，我们展示的是最大模块度M 的划分。图a和图c旁边给出了相应划分的归一化互信息和社区数n_c 。归一化互信息（公式**8.23** ）是针对非重叠社区提出的，也可以应用到重叠社区 [59] 。

LFR基准网络（Lancichinetti-Fortunato-Radicchi）

GN基准网络生成一个随机图，所有节点的度相近，所有社区的大小相等。然而，绝大多数真实网络的度分布是重尾分布，社区大小分布也是重尾分布（图**8-29** ）。因此，在GN基准网络上性能良好的算法未必在真实网络上也有良好的性能。为了避免这一局限性，LFR基准网络（图**8-26** ）生成的网络的度分布和预设社区大小分布都服从幂律 [58] 。

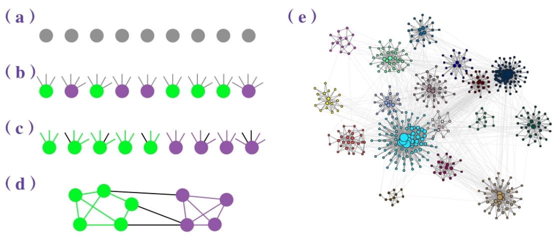


图**8-26**　**LFR**基准网络

在LFR基准网络的构造过程中，生成的网络的节点度和社区大小都服从幂律。构造过程如下 [57] ：

（**a**） 从N 个不连通节点开始。

（b） 给每个节点赋予一个社区，该社区有N_c 个节点，N_c 服从社区指数为ζ 的幂律分布$P_{N_c} \sim N_c^{-\zeta}$ ，同时也从度指数为γ 的幂律分布$P_k \sim k^{-\gamma}$ 中生成度k_i 赋给每个节点i 。

（c） 一个社区的每个节点i 获得内部度$(1-\mu)k_i$ ，表示为来自相同颜色的链接。余下μk_i 的度以黑色链接表示，连接到其他社区的节点。

（d） 同一社区的节点的所有半链接随机彼此相连，直到不再有“自由”的半链接。这样我们保持了社区中每一个节点的内部度的序列。余下的μk_i 个半链接随机连接到其他社区的节点。

（e）$N=500$、$\gamma=2.5$且$\zeta=2$的LFR基准网络生成的典型网络及其社区结构。

建好了已知社区结构的网络，接下来我们需要找到工具，来度量每种社区识别算法给出的划分的准确率。在此期间，我们需要牢记前述两个基准网络对应一种特定的社区定义。而基于团渗流或链接聚类的算法，在这两个网络上可能不会有很好的表现，因为它们是针对其他社区定义提出的。

度量准确率

为了比较识别的社区和基准网络中预设的社区，考虑任意一个非重叠社区的划分。在每一步，我们随机选择一个节点，记录它所属的社区标签。我们会得到由服从$p(C)$分布的社区标签构成的一个随机字符串，其中$p(C)$表示随机选择的节点属于社区C 的概率。

考虑同一网络的两种划分，其中一个是基准网络的社区划分（标准答案），另一个是由某种社区识别算法找到的社区。每种划分有自己的$p(C_1)$和$p(C_2)$分布。联合分布$p(C_1, C_2)$是一个随机选择的节点在第一种划分里属于社区C_1 且在第二种划分里属于C_2 的概率。两种划分的相似度可以由归一化互信息表示[38] ：

$$I_n = \frac{\sum_{C_1,C_2} p(C_1,C_2)\log_2 \frac{p(C_1,C_2)}{p(C_1)p(C_2)}}{\frac{1}{2}H\left(\{p(C_1)\}\right) + \frac{1}{2}H\left(\{p(C_2)\}\right)} \qquad (8.19)$$

公式8.19 的分子是互信息I ，I 可以度量两种社区划分共享的信息：若C_1 和C_2 相互独立，则$I=0$；若两种划分完全相同，则当

$H\left(\left\{p\left(C_1\right)\right\}\right)=H\left(\left\{p\left(C_2\right)\right\}\right)$ 时，I 取得最大值，其中：

$$H\left(\left\{p(C)\right\}\right)=-\sum_{C}p(C)\log_2 p(C) \tag{8.20}$$

公式**8.20** 是香农熵。

如果所有节点都属于同一社区，那么我们很清楚下一个随机选择的节点所属的社区，即$H=0$，也就是说，检查下一个节点属于哪个社区并不会给我们带来新的信息。当p（C）是均匀分布时，H 最大，此时我们不知道下一个节点属于哪个社区，每个新节点都会带来H 比特的新信息。

总而言之，如果基准网络和发现的划分是相同的，则$I_n=1$，如果它们彼此独立则$I_n=0$。I_n 的用处如图**8-25b** 所示，展示了劳沃斯算法在GN基准网络上的准确率。图**8-27** 展示了在GN和LFR基准网络上使用I_n 检验每种算法性能的结果。根据检验结果，我们得出以下结论：

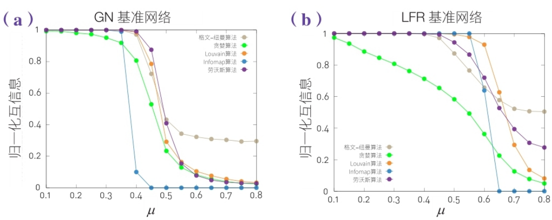


图**8-27**　使用基准网络检验

我们在GN和LFR基准网络上检验了识别非重叠社区的每种社区识别算法。图中展示了5种算法的归一化互信息I_n 关于μ 的函数。每种算法的名称参见表**8-1** 。

（**a**）**GN**基准网络

横轴为混合参数（公式**8.18**），表示连接不同社区的链接的比例。纵轴是归一化互信息（公式**8.19**）。每条曲线都是100次独立实现的均值。

（**b**）**LFR**基准网络

LFR基准网络上的与（a）相同的实验。基准网络参数为$N=1\ 000$，$\langle k\rangle=20$，$\gamma=2$，$k_{\max}=50$，$\zeta=1$，最大社区：100个节点，最小社区：20个节点。每条曲线都是25次独立实现的均值。

当$\mu <0.5$时，有$I_n \simeq 1$。因此，当社区内链接密度相对高于社区外时，大部分算法能准确地识别预设的社区。高于$\mu =0.5$时，每种算法的准确率都会下降。

准确率与基准网络选择有关。在更接近真实情况的LFR基准网络上，Louvain算法和劳沃斯算法的性能较好，贪婪模块度算法的性能较差。

速度

如**8.2**节 所讨论的，可能的划分数随N 增长的速度比指数还快。这使得大多数真实网络的可能划分数是天文数字。虽然社区识别算法不检查所有可能的划分，但其计算开销相差很大，影响了算法的速度和可以处理的网络大小。表**8-1** 总结了本章讨论的算法的计算复杂度。因此，最高效的算法是Louvain算法和Infomap算法，两者的复杂度均为$O(N \log N)$。最低效的算法是团渗流算法，其复杂度为$O(e^N)$。

表**8-1** 不同算法的计算复杂度

算法名	实质	计算复杂度	参考文献
劳沃斯算法	层次凝聚	$O(N^2)$	[11]
格文－纽曼算法	层次分裂	$O(N^3)$	[9]
贪婪算法	最优化模块度	$O(N^2)$	[33]
优化的贪婪算法	最优化模块度	$O(N\log^2 N)$	[35]
Louvain 算法	最优化模块度	$O(L)$	[2]
Infomap 算法	最优化流	$O(N\log N)$	[44]
团渗流算法	重叠社区	$\exp(N)$	[48]
链接聚类算法	层次凝聚；重叠社区	$O(N^2)$	[51]

本章讨论的社区识别算法的计算复杂度。虽然计算复杂度依赖于N 和L 二者，但对于稀疏网络，我们完全可以近似地认为$L \sim N$ 。因此，我们列出的计算复杂度只关于N 。

然而这些复杂度表达式并不表示真实的运行时间。它们只表明了运行时间如何随N 变化。在需要对非常大的网络识别社区时，复杂度的量级很重要。为了了解这些算法的真实速度，我们在同一台计算机上测量了它们在蛋白质相互作用网络（$N =2\ 018$）、电网（$N =23\ 133$）和科学合作网络（$N =4\ 941$）上的运行时间。结果如图**8-28**

所示，表明：

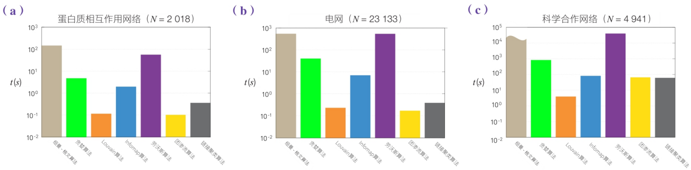


图8-28　运行时间

我们使用各个算法的Python程序来比较社区识别算法的速度，要么基于其开发者发布的版本，要么基于igraph软件包里的实现。我们自己实现了劳沃斯算法，因为它未经优化，运行时间可能长于理想状态。每个算法都在同一台计算机上测试。图中给出了三个真实网络上的运行时间，单位是秒。科学合作网络上的格文－纽曼算法在7天后仍未完成，因此我们只能给出它的运行时间的下限。科学合作网络的运行时间较长，是因为该网络较大。

- Louvain算法在所有网络上所需的运行时间最短。团渗流算法只在中等规模的网络上速度快，在较大的合作网络上，其运行时间与其他算法差不多。
- 格文－纽曼算法在每个网络上都是最慢的，这与其计算复杂度高是一致的（表**8-1**）。例如，该算法未能在7天内找到科学合作网络上的社区。

总而言之，基准网络让我们得以比较可用算法的准确率和速度。设计出最快和最准确的算法依然是一场活跃的军备竞赛，对这个问题感兴趣的人应该参考在多个维度上比较这些算法的文献 [31] ，[58] ，[60] ，[61] 。

8.7
刻画社区特征

网络科学研究的驱动力是量化探索网络的基础原理，这些基础原理决定了网络是如何生成和组织的。组织原理影响了社区结构和我们识别社区的能力。本节将讨论网络演化、社区大小分布的特征、链接权重在社区识别中的作用，从而揭示社区组织的一般性原理。

社区大小分布

根据基础假设（H1），网络中社区的数量和大小都由网络的连接模式唯一确定。因此，我们不禁要问：社区大小是如何分布的？

许多研究表明，社区大小服从重尾分布，同时存在着大量小社区和少数几个超大社区 [16]，[33]，[35]，[36]，[60]。为了探索重尾分布是否普遍存在，我们在图**8-29** 中画出三个网络上由不同社区识别算法给出的社区的p_{N_c}。图形中可以看到如下模式：

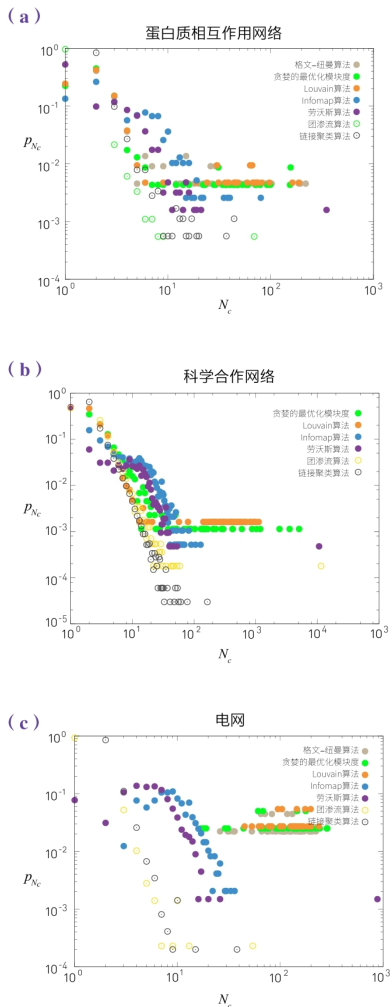


图8-29　社区大小分布

本章讨论的社区识别算法给出的社区大小分布p_{N_c}。算法名称参见表**8-1**。在蛋白质相互作用（图a）和科学合作网络（图b）上，所有算法都给出近似重尾的社区大小分布，因此这些结果多多少少相互一致。如图c所示，不同算法在电网上给出了不一致的结果。

- 在蛋白质相互作用网络和科学合作网络上，所有算法都给出近似重尾分布的p_{N_c}。因此，在这两个网络上同时存在着为数众多的小型社区和少数大型社区。
- 电网上不同的算法给出不同的结果。基于模块度的算法给出的社区的大小相近，约为$N_c \simeq 10^2$。相比而言，劳沃斯算法和Infomap算法给出大量的$N_c \simeq 10$的小社区和少数大型社区。最后，团渗流和链接聚类给出近似重尾的社区大小分布。

上述差异表明，社区大小的重尾分布并不是某一种算法的副产品，而是诸如蛋白质和科学合作网络等某些网络的固有属性。电网上的不同结果表明，该网络并不具有唯一的可识别的社区结构。

社区和链接权重

链接权重与社区结构密切相关。然而，正如我们接下来要讨论的，这一相关性本质上是依赖于系统本身的。

社交网络

两个人在一起的时间越多，他们越可能有共同的朋友，这使得他们属于同一个社区的概率也变大了。因此，社交网络的社区倾向于以强关系为核心。相比而言，不同社区之间的链接较弱。图**8-30a**所示的手机通话网络很好地展示了这种模式[63]。这种模式被称为弱链接假设[62]。我们可以看到，强链接确实主要出现在大量小社区内部，而社区之间的链接较弱。

交通系统

许多技术和生物网络的用途是传输物质或信息的。在这种情况下，链接权重应该与介数中心度有关 [64-66] 。后者表征了网络中的局部交通流。因为社区之间的链接需要传输相当数量的物质或信息，交通网络中社区之间的链接是强链接。相比而言，社区内部的连接较弱（图**8-30b** ）。

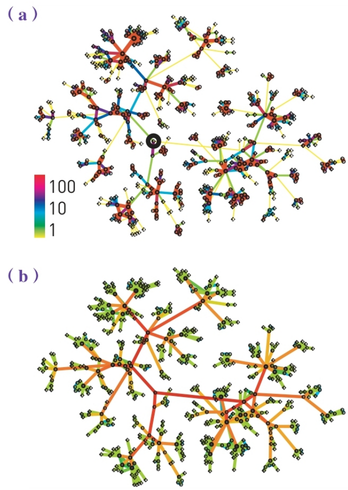


图**8-30**　社区和链接权重

手机通话网络展示了链接权重和社区之间的关系。链接表示用户之间的相互通话。这里只显示距离图a中黑色圆圈高亮的用户6步或6步以内的节点。

（**a**）真实权值

链接颜色表示通话总时长的分钟数（见颜色图例）。与弱链接假设一致，我们发现强链接主要出现在社区内，弱链接主要出现在社区之间 [62] 。

（**b**）介数中心度

如果链接权重像技术和生物系统中那样表示信息或物质传输的需求，权值则可以用介数中心度良好近似（图**8-11** ）。我们用颜色来标示每条链接的介数中心度。如图所示，社区之间的链接的中心度较高（红色），而社区内部的链接的中心度较低（绿色）。 [63]

链接权重和社区结构的这一耦合性质表明，引入链接权重可能会提高社区识别算法的准确率。然而，社交系统和技术系统的差异特质提醒我们：旨在把强链接的节点放入同一社区的算法很可能只对社交系统有效，而在社区之间强链接的技术和生物系统中可能会给出误导

性结果。

社区演化

网络连接模式的改变对社区有多重作用：可能促使新社区出现，已有社区增长或缩小，社区合并或分裂成几个小社区，甚至社区消亡（图**8-31**）[50]。针对社交网络和通信网络的研究给出了关于社区改变的一些看法[50]，[67-73]：

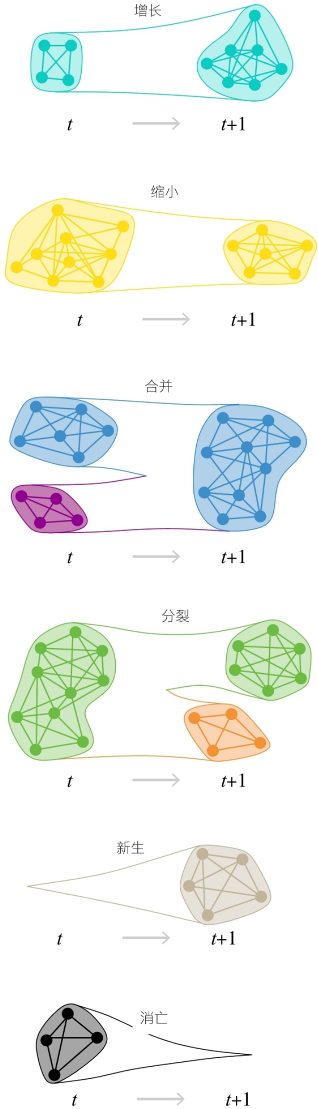


图8-31　演化社区

当网络随时间演化，底层的社区结构也随之演化。社区结构的任何改变都是社区生存周期中这6种基本事件的结果，如图所示：社区可以增长或缩小，合并或分裂，新生或消亡。[50]

增长

随着一个节点与某个社区的成员之间的链接增多，该节点加入该社区的概率也随之增大[73]。

缩小

相比于有很多链接连接到社区内部其他成员的节点，那些只有少数几条链接连接到社区内部其他成员的节点更容易离开社区[73]。在加权网络上，节点离开社区的概率随着其指向社区外的链接权重之和增大而增大。

分裂或消亡

社区解体的概率随着指向社区外的链接权重之和增大而增加。

年龄

社区的年龄和大小正相关，表明老社区更可能是大社区[50]。

社区稳定性

大社区的成员随时间变化得比小社区的成员快。事实上，在社交网络中，大社区通常对应机关、公司或学校，会接收新成员、雇用新员工或招收新学生。小社区的稳定性要求成员稳定[50]。

上述结论是在社交系统中得出的。我们对于技术或生物系统的社区演化的模式目前所知甚少。

总而言之，几个重复出现的模式刻画了社区的组织形式与演化。社区大小分布通常是重尾的，表明同时存在着大量小社区和少数大社区。我们也发现社区结构与链接权重之间有关联，且这种关联与系统相关：社交系统中强链接通常在社区内，而交通系统中强链接通常在社区之间。最后，我们对决定社区演变的动态模式有了更深入的理解。

8.8

小结

社区在不同网络中普遍存在，这使得社区识别成为网络科学领域的一个热门话题。目前提出的算法大多提供软件包以便随时用于网络分析。然而，算法的易用性和对其结果的解释要求我们关注算法基于的假设。本章给出了社区识别的基本知识和量化方法，以便我们理解这些常用算法的起源和背后的假设。

尽管社区识别取得了成功，但该领域依然面临许多未解决的问题：

社区真的存在吗？

本章我们回避了一个基础问题：如何判断一个网络中真的存在社区？换言之，我们能否在识别社区之前判断一个网络有没有社区？这个尚无答案的问题或许是社区识别领域最明显的空白。社区识别算法是用来识别（无论是否存在）社区的。

假设还是定理？

如边栏**8.3** 所总结的，社区识别依赖于4个假设，之所以称之为假设是因为我们没法验证它们正确与否。未来的研究也许可以将基础假设、随机假设和最大模块度假设变为定理。或许，我们也会像认识到最大模块度假设的局限性（**8.6**节 ）一样认识到其他假设的局限性。

边栏**8.3**

快速回顾：社区

社区识别依赖于4个有关社区本质的假设。

基础假设

网络的社区结构仅由其连接模式决定。这个标准答案是隐藏

的，我们需要设计合适的算法来发现。

连通性和密度假设

社区是网络中局部紧密连接的子图。

随机假设

随机连接的网络没有内在社区结构。

最大模块度假设

给定一个网络，模块度最大的划分对应最优的社区结构。其中模块度由下式定义：

$$M=\sum_{c=1}^{n_c}\left[\frac{l_c}{L}-\left(\frac{k_c}{2L}\right)^2\right]$$

所有节点都必须属于社区吗？

社区识别算法要求所有节点都属于社区。对大多数真实网络来说，这个要求可能太严格了：有些节点属于单个社区，有些属于多个社区，可能很多节点不属于任何社区。社区识别的绝大多数算法没有对此做出区分，而是要求所有节点都属于某个社区。

密集社区和稀疏社区

本书讨论的绝大多数网络都是稀疏的。然而，随着数据采集的进步，可以获取更多真实的链接数据。在密集网络中，我们通常会看到高度重叠的社区。这要求我们重新审视本章讨论的各个假设是否有效，各种算法是否适用。例如，在高度重叠的社区里，节点的外部度可能高于内部度，这使得密集假设的有效性存疑。

社区是否有意义？

我们用一个例子来回答这个问题。图**8-32a** 展示了手机通话网络

的局部邻域，并高亮了链接聚类算法识别的4个社区（**8.5**节 ）。图中也展示了正午**8-32b** 和午夜**8-32c** 的通话频率，以比较人们在一天中不同时候的不同通话习惯。我们发现右上角社区的成员，即**8-32a** 中的棕色节点，在午夜很活跃**8-32b** ，但正午不打电话**8-32c** 。与之相比，浅蓝色和深蓝色社区在正午活跃，在午夜时休息。这表明，虽然社区的识别是完全基于网络连接模式的，但依然存在一致的社区活动模式。

图**8-32** 表明，一旦存在社区，社区会对网络行为产生深刻影响。大量实践观测支持这一结论：信息在社区内部传播得快，但很难到达其他社区[63] ；社区影响链接权重[62] ；社区的存在会致使度相关性[74] 。

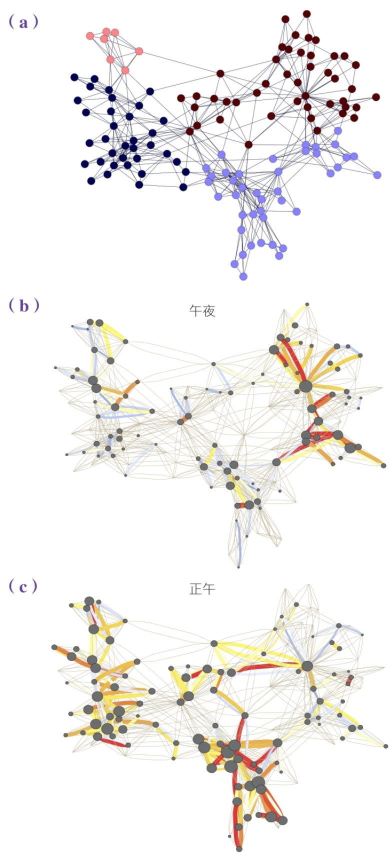


图8.32　社区和通话模式

手机通话网络既包含社区结构，也包含用户活动的信息，展示了社区对成员活动的直接影响。

（**a**）社区结构

手机网络的4个社区。每个社区以不同颜色表示。链接聚类算法给出的这些社区表示了超过100万消费者的通话模式的局部邻域（**8.5**节　　）。手机网络的其余部分未画出。

（**b**）午夜活跃程度

图a中所示4个社区用户的通话模式。链接颜色表示午夜时段的通话频率。红色链接表示午夜前后有很多通话，白色或缺失链接表示用户在午夜时段很少通话或者完全不通话。

（**c**）正午活跃程度

与图b类似，但表示正午的活跃程度。

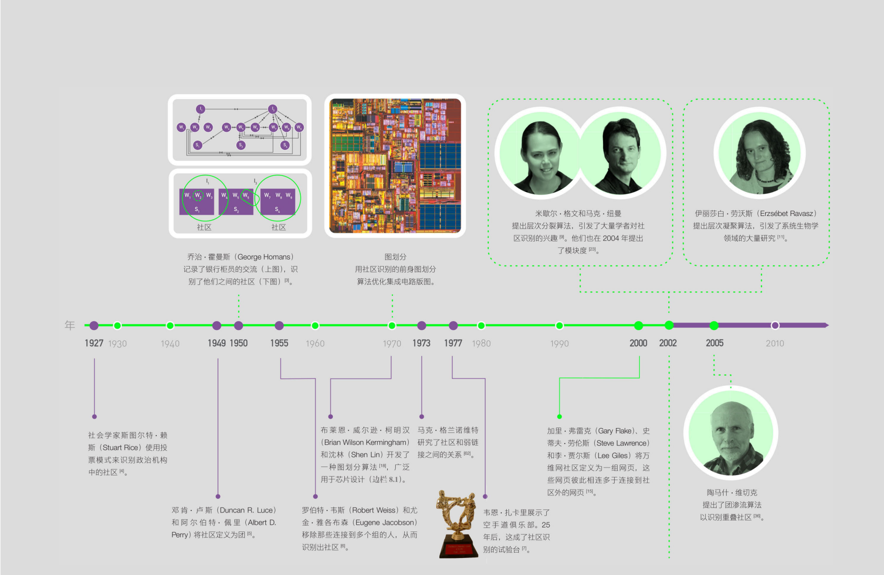
感谢苏内·莱曼（Sune Lehmann）提供图片。

社区在潜在应用方面同样重要。例如，让万维网社区内部的各个客户端之间的链接增强，可以提高基于网络的服务的性能[75]。在市场营销中，社区识别可以用于找到相似兴趣或购买习惯的消费者，帮助设计高效的产品推荐系统[76]。社区常被用于制定数据结构以便快速处理数据查询[77]，[78]。最后，Facebook、Twitter、领英等许多社交网站背后的社区识别算法帮助寻找潜在的朋友、感兴趣的职位和广告目标。

虽然社区识别植根于社会科学和计算机科学，但它在网络科学中是很年轻的领域（边栏**8.4**　）。因此，我们对社区组织的认识仍在快速发展，以创造出更准确的工具来分析大型网络中的局部结构。

边栏8.4

社区识别简史



8.9
课后习题

习题1　层次网络

计算图8-33 所示层次网络的度指数。

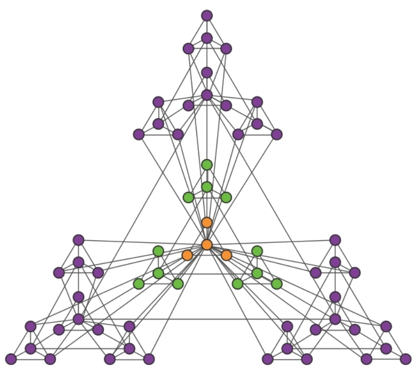


图8-33　层次网络

不同颜色表示网络构造过程的不同阶段。

习题2　环上的社区

考虑有N 个节点的一维格子构成的一个环，其中每个节点连接到两个邻居。将这个环划分成n_c 个连续的簇，每个簇的大小为$N_c = N / n_c$ 。

（1）计算这个划分的模块度。

（2）根据最大模块度假设（**8.4**节），M_c 的最大值对应最优划分。找到最优划分的社区大小n_c 。

习题3　模块度分辨率极限

考虑一个网络，由n_c 个团构成环状，每个团有N_c 个节点和$m(m-1)/2$条链接。相邻团之间通过一条链接相连（图**8-34**）。网络有明确的社区结构，每个社区对应一个团。

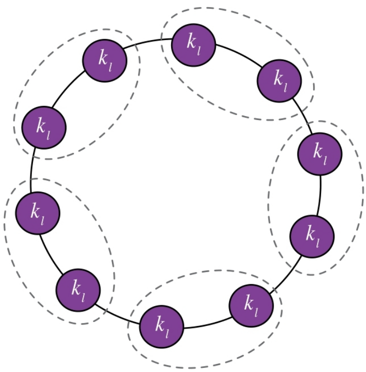


图**8-34**　模块度

（1）计算这一自然划分的模块度M_{single} ，以及两两配对的划分的模块度M_{pairs} 。后者将相邻的团合并为一个社区，如图**8-34**虚线所示。

（2）证明只有当$n_c < 2L$ 时，最大模块度才能给出符合直观的正确社区划分，其中：

$$L = \frac{n_c m(m-1)}{2} + n_c$$

（3）讨论上述不等式不成立的情况。

习题4　最大模块度

证明公式8.12 定义的模块度M 的最大值不会超过1。

8.10 进阶阅读8.A 层次模块度

本节我们讨论图8-13　介绍的层次模型的标度属性。我们计算度分布以及依赖于度的集聚系数，从而推导出公式8.8　。最后，我们在10个真实网络上讨论层次结构是否存在。

度分布

为了统计模型的度分布，我们计算各种度的节点个数。从图**8-13a**　的第一个模块的5个节点开始，我们把中心节点标记为枢纽节点，其余4个为外围节点。复制后的中心节点仍然称为枢纽节点，复制后的外围节点仍然称为外围节点（图**8-35**　）。

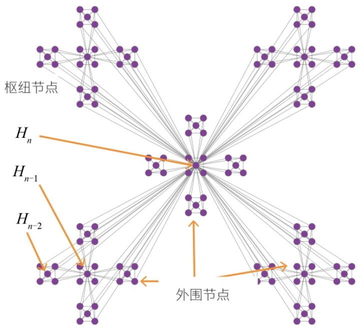


图8-35　计算度指数

层次网络的结构和枢纽节点的命名约定。[11]

网络中心最大的枢纽节点在第n 轮迭代中得到4^n 条链接。我们称这个中心枢纽节点为H_n ，它的4个复制品为H_{n-1} （图**8-35** ）。我们将余下的4×5个模块中心称为H_{n-2} ，这些模块的大小等于网络在第n -2轮的大小。

在第n 轮迭代时，枢纽节点H_i 的度服从公式：

$$k_n(H_i)=\sum_{l=1}^{i}4^l=\frac{4}{3}(4^i-1) \quad (8.21)$$

其中，我们使用以下公式：

$$\sum_{l=0}^{i}x^l=\frac{x^{i+1}-1}{x-1} \quad (8.22)$$

或者公式：

$$\sum_{l=1}^{i}x^l=\frac{x^{i+1}-1}{x-1}-1 \quad (8.23)$$

当i < n 时，H_i 模块的个数为：

$$N_n(H_i)=4\times 5^{n-i-1} \quad (8.24)$$

也就是说，$i=n$ -1有4个模块，$i=n$ -2有4×5个模块……，$i=n$ 有$4\times 5^{n-2}$ 个模块。因为我们有$4\times 5^{n-i-1}$ 个H_i 枢纽节点，它们的度为$k_n(H_i)$，则根据公式**8.21** 和**8.24** 可以得出：

$$\ln N_n(H_i)=C_n-i\cdot\ln 5 \quad (8.25)$$

$$\ln k_n(H_i)\backsimeq i\cdot\ln 4+\ln(4/3) \quad (8.26)$$

其中C_n 为：

$$C_n=\ln 4+(n-1)\ln 5 \quad (8.27)$$

在公式**8.26** 中我们近似认为$4^i-1\backsimeq 4i$ 。

对于所有的$k>n+2$，我们可以合并公式**8.26** 和公式**8.27** ，得

到：

$$\ln N_n(H_i) = C_n^{'} - \ln k_i \frac{\ln 5}{\ln 4} \tag{8.28}$$

或者得到：

$$N_n(H_i) \sim k_i^{-\frac{\ln 5}{\ln 4}} \tag{8.29}$$

若想计算度分布，我们需要计算以下比率来归一化$N_n(H_i)$：

$$p_{k_i} \sim \frac{N_n(H_i)}{k_{i+1} - k_i} \sim k_i^{-\gamma} \tag{8.30}$$

根据以下公式：

$$k_{i+1} - k_i = \sum_{l=1}^{i+1} 4^l - \sum_{l=1}^{i} 4^l = 4^{i+1} = 3k_i + 4 \tag{8.31}$$

可以得到：

$$p_{k_i} = \frac{k_i^{-\frac{\ln 5}{\ln 4}}}{3k_i + 4} \sim k_i^{-1-\frac{\ln 5}{\ln 4}} \tag{8.32}$$

换言之，该层次网络的度指数为：

$$\gamma = 1 + \frac{\ln 5}{\ln 4} = 2.16 \tag{8.33}$$

集聚系数

计算H_i 枢纽节点的集聚系数非常简单。节点的$\sum_{l=1}^{i} 4^l$ 条链接来自各个正方形的顶点，而这些顶点之间的连接数等于它们的个数。因此，H_i 的邻居之间的链接数为：

$$\sum_{l=1}^{i} 4^l = k_n(H_i) \tag{8.34}$$

由此可以得出：

$$C(H_i)=\frac{2k_i}{k_i(k_i-1)}=\frac{2}{k_i-1} \quad (8.35)$$

换言之，我们可得：

$$C(k)\simeq\frac{2}{k} \quad (8.36)$$

表明枢纽节点的$C(k)$量级为k^{-1}，这与公式**8.8** 一致。

实证结果

图**8-36** 展示了10个参照网络的$C(k)$函数。我们同时展示了每个网络保度随机化之后的$C(k)$（绿色记号），并观测到了以下现象：

- k 较小时，所有网络的$C(k)$都比随机化网络的$C(k)$高一个量级。因此小度节点所处的邻域的密集程度高于随机期望。
- 对于科学合作、代谢和引文网络，我们可以良好地近似为$C(k)\sim k^{-1}$，而随机化网络的$C(k)$函数图像是水平线。因此，这些网络表现出的图**8-13** 模型的层次模块度。
- 对于互联网、手机通话网络、演员网络、电子邮件网络、蛋白质相互作用网络和万维网，随机化网络的$C(k)$与k 无关。因此这些网络表现出层次模块度，观测到的$C(k)$无法由我们的简单层次模型来刻画。要拟合这些系统的$C(k)$，我们需要构造模型来准确刻画其演化过程。这类模型给出$C(k)\sim k^{-\beta}$，其中β 可以不等于1 [27]。
- 我们只在电网上观测到$C(k)$是水平线，即与k 无关。这表明该网络不存在层次模块度。

综合起来，图**8-36** 表明大多数真实网络有非同小可的层次模块

度。

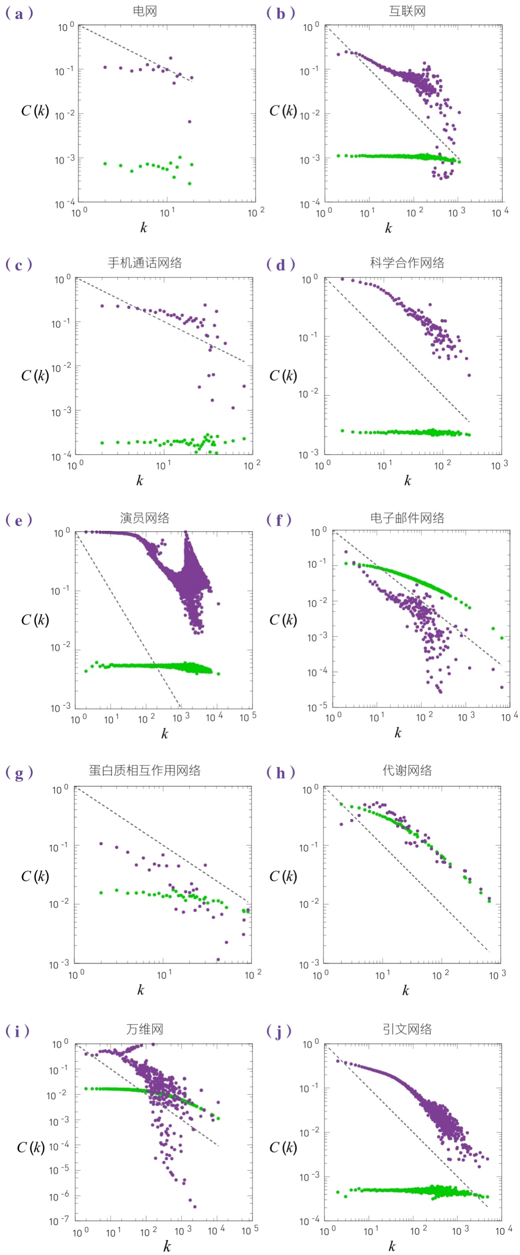


图8-36　真实网络的层次结构

10个参照网络上$C(k)$与k 的关系（紫色记号）。对每个网络做保度随机化之后的$C(k)$以绿色记号表示。保度随机化抹平了局部密度的差异，因此社区和底层的层次结构都会消失。有向网络被转换成无向网络以便度量$C(k)$。每个图中的短虚线的斜率为-1，参照公式**8.8**，用于指引读者观察。

8.11
进阶阅读8.B
模块度

本节我们推导公式**8.12** 和公式**8.13**，即模块度函数及其变化量。

模块度改写成在社区上求和的函数

根据公式**8.9** 和公式**8.10**，我们可以得出整个网络的模块度：

$$M = \frac{1}{2L}\sum_{i,j=1}^{N}(A_{ij} - \frac{k_i k_j}{2L})\delta_{C_i,C_j} \quad (8.37)$$

其中C_i 是节点i 所属社区的标签。只有同属一个社区的节点对才会在公式**8.37** 的求和中有贡献，我们把第一项重写为对社区求和：

$$\frac{1}{2L}\sum_{i,j=1}^{N} A_{ij}\delta_{C_i,C_j} = \sum_{c=1}^{n_c}\frac{1}{2L}\sum_{i,j\in C_c} A_{ij} = \sum_{c=1}^{n_c}\frac{L_c}{L} \quad (8.38)$$

其中L_c 是社区C_c 的内部链接数。因子2被消掉是因为A_{ij} 中每条链接算了两遍。

与之类似，公式**8.37** 的第二项变成：

$$\frac{1}{2L}\sum_{i,j=1}^{N}\frac{k_i k_j}{2L}\delta_{C_i,C_j} = \sum_{c=1}^{n_c}\frac{1}{(2L)^2}\sum_{i,j\in C_c} k_i k_j$$

$$= \sum_{c=1}^{n_c}\frac{k_c^2}{4L^2} \quad (8.39)$$

其中k_c 是社区C_c 的节点度之和。事实上，在配置模型中，半链接

连接到另一个随机选择的半链接的概率是 $\frac{1}{2L}$ ，因为整个网络中的半链接总数为$2L$ 。因此，我们的半链接连接到模块内的另一个半链接的可能性为 $\frac{k_c}{2L}$ 。对社区C_c 内所有k_c 个半链接重复这一过程，加上1/2以避免重复计数，可得到公式**8.39** 的最后一项。

结合公式**8.38** 和公式**8.39** 可得公式**8.12** 。

合并两个社区

考虑社区A和B，以k_A 和k_B 表示两个社区各自的节点度之和（等价于上文的k_C ）。我们希望计算合并这两个社区之后模块度的变化量。根据公式**8.12** ，变化量可以写作：

$$\Delta M_{AB}=\left[\frac{L_{AB}}{L}-\left(\frac{k_{AB}}{2L}\right)^2\right]-\left[\frac{L_A}{L}-\left(\frac{k_A}{2L}\right)^2+\frac{L_B}{L}-\left(\frac{k_B}{2L}\right)^2\right] \quad (8.40)$$

$$L_{AB}=L_A+L_B+l_{AB} \quad (8.41)$$

l_{AB} 是社区A的节点和社区B的节点之间的链接数。且有：

$$k_{AB}=k_A+k_B \quad (8.42)$$

将公式**8.41** 和公式**8.42** 代入公式**8.40** 可得：

$$\Delta M_{AB}=\frac{l_{AB}}{L}-\frac{k_Ak_B}{2L^2} \quad (8.43)$$

即为公式**8.13** 。

8.12
进阶阅读8.C
社区识别的快速算法

本章选择讨论的算法旨在展示与社区识别相关的基本思想和概

念。因此这些算法并不一定是最快的，也不一定是最准确的。近来有两种算法被普遍应用，分别为Louvain算法和Infomap算法。它们的准确率与本章讨论的算法的准确率相当，但可扩展性更好。因此我们可以使用这两种算法在大型网络上识别社区。

这两种算法有诸多相似之处：

- 两种算法的目标都是优化某个Q 。对于Louvain算法，Q 是模块度M 。对于Infomap算法，Q 是基于熵的一种度量，被称为映射方程或L 。
- 两种算法都使用同样的优化过程。

基于上述相似性，我们把这两种算法放在一起讨论。

Louvain算法

计算复杂度$O(N^2)$的贪婪算法可能无法用于非常大的网络。布隆代尔和合作者提出了一种可扩展性更好的模块度优化算法 [2] 。Louvain算法由重复迭代的两个步骤组成（图**8-37** ）。

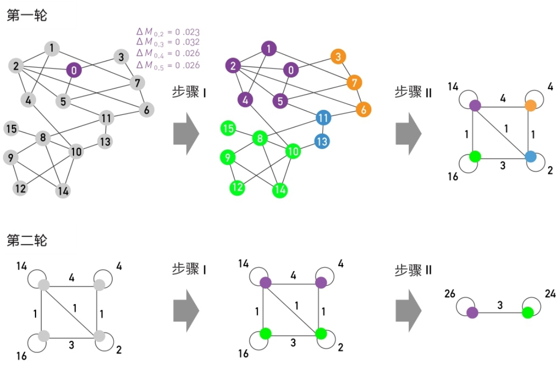


图8-37　Louvain算法

Louvain算法的主要步骤。每一轮由两个步骤组成。

步骤I

通过局部改变来优化模块度。我们选择一个节点，并计算若该节点加入邻居社区导致的模块度变化量（公式**8.44** ）。图中展示了节点0的期望模块度变化量$\Delta M_{0,i}$ 。则

节点0应并入节点3，因为这模块度的变化量最大，为$\Delta M_{0,3}=0.032$。对每个节点重复这一过程，最后的社区标记如节点颜色所示。步骤I结束。

步骤II

整合步骤I所得的社区，建立一个新的社区网络。属于同一社区的节点都合并为一个节点，如右上角所示。这一过程会产生自环，表示合并为一个节点的社区内部节点之间的链接。

步骤I和步骤II合起来为一轮。对每一轮所得的网络再次进行同样处理（第二轮），直到模块度不可能再增加为止。[2]

步骤I

从一个有N 个节点的加权网络开始，每个节点都属于不同的社区。把每个节点i 依次放入所有邻居j 所属的社区中，分别计算模块度的增益。如果最大的增益为正，则将i 移入最大增益对应的社区。如果没有增益为正，则i 留在原来的社区。对所有节点重复一遍这个过程，直到模块度无法再增加，此时步骤I完成。

把一个不连通节点i 移入社区C所导致的模块度变化量ΔM 为：

$$\Delta M=\left[\frac{\sum_{\text{in}}+2k_{i,\text{in}}}{2W}-\left(\frac{\sum_{\text{tot}}+k_{\text{i}}}{2W}\right)^2\right]-\left[\frac{\sum_{\text{in}}}{2W}-\left(\frac{\sum_{\text{tot}}}{2W}\right)^2-\left(\frac{k_i}{2W}\right)^2\right] \tag{8.44}$$

其中Σ_{in} 是C内部的链接权重之和（无向图的L_c ）；Σ_{tot} 是C的所有节点的链接权重之和；k_i 是连入节点i 的链接权重之和；$k_{i,\text{in}}$ 是节点i连接到C内部节点的链接权重之和；W 是网络中所有链接权重之和。

我们注意到ΔM 是公式**8.13** 的特例，后者是合并两个社区A和B之后模块度的变化量。本例中B是一个不连通节点。我们也可以用ΔM来确定将i 从原社区中移除的模块度变化量，只需计算将i 合并到（已排除掉i 的）社区C的ΔM ，则移除i 的改变量即为-ΔM 。

步骤II

我们构造一个新网络，其节点是步骤I识别的社区。两个节点之间的链接权重是对应社区的节点之间的链接权重之和。社区内的节点之间的链接表示为加权自环。

步骤II完成后，重复步骤I和步骤II。我们将这样两个步骤合称为一轮（图**8-37**）。社区数每轮都会减少。重复操作直到模块度不再有变化，我们就找到了最大模块度。

计算复杂度

Louvain算法更多地受到存储需求而非计算时间的限制。算法的第一轮迭代耗时最多，计算次数随L呈线形关系。因为后续各轮的节点数和链接数都减少了，算法的复杂度最多为$O(L)$。因此我们可以在数百万节点的网络上识别社区。

Infomap算法

Infomap算法由马丁·罗斯瓦尔（Martin Rosvall）和卡尔·伯格斯特龙（Carl T. Bergstrom）提出，这种算法利用数据压缩来识别社区（图**8-38**）[44-46]，方式是在有向有权网络上优化一个社区识别品质函数，被称为映射方程（map equation）。

考虑一个网络被划分成了n_c个社区。我们希望以最高效的方式编码该网络上一个随机游走者的轨迹。换言之，我们希望以最少个数的符号来描述轨迹。理想的编码应该利用这一事实：随机游走者往往被困在社区内很长时间（图**8-38c**）。

为了找到这样一种编码，我们分配给：

- 每个社区一个代码（索引代码本）。例如图**8-38c**的紫色社区被分配了代码111。
- 每个社区的每个节点一个代码字。例如图**8-38c**中左上角节点被分配了001。我们注意到一个节点代码可以在不同社区中反复使用。
- 退出码表示随机游走者离开了一个社区，例如图**8-38c**

中紫色社区的0001。

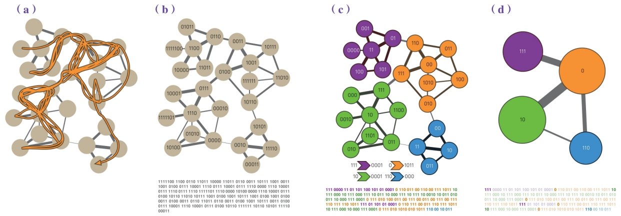


图8-38　从数据压缩到社区

Infomap算法通过压缩网络上随机游走者的移动来识别社区。

（a）　橙色实线表示一个小网络上随机游走者的轨迹。我们希望以最少的符号描述这条轨迹。为了达到这个目标，我们给重复访问的结构（社区）分配唯一的短名字。

（b）　一开始，我们根据霍夫曼编码给每个节点分配唯一的名字。霍夫曼编码是一个数据压缩算法，根据随机游走访问每个节点的估计概率来给节点分配代码。网络下方的314个比特描述了图a中所示的随机游走者的轨迹，从左上角第一个游走节点1111100开始，然后是第二个节点1100，以此类推，最终停止在右下角的最后一个游走节点00011。

（c）　图中展示了随机游走的二级编码，每个社区有唯一的名字，社区内的节点名可以复用。这种编码平均减少了32%的编码长度。社区名字代码和退出每个社区的代码分别标注在图的左侧和网络下方箭头右侧。使用这种编码，我们可以用图c的网络下方给出的243个比特来描述图a。前3个比特111表示游走从红色社区开始，代码0000表示第一个游走节点，以此类推。

（d）　我们只给出了社区名字，忽略社区内各个节点的位置，就得到了网络的高效粗粒度表达，对应社区结构。

因此，我们的目标是建立一套编码，对随机游走提供最短描述。一旦找到这套编码，我们就可以从索引代码本中识别出社区——索引代码本给每个社区都分配了唯一编码（图**8-38c**）。

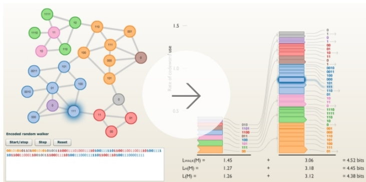
寻找映射方程的最小值以找到最优编码：

$$\mathcal{L} = qH(Q) + \sum_{c=1}^{n_c} p_{\circlearrowright}^{c} H(P_c) \qquad (8.45)$$

简而言之，公式**8.45**　的第一项给出了描述社区间移动所需的平均比特数，其中q　是随机游走者在某一步切换社区的概率。第二项给

出了描述社区内移动所需的平均比特数。这里H（P_c）是社区内移动的熵，包括离开社区i的退出码。

映射方程的具体表达和基于随机游走者在网络上移动概率的计算是相关联的[44-46]。在线资源**8.3** 提供了一个交互式工具，以展示公式**8.45** 背后的机制及其应用。



在线资源**8.3**　**Infomap**算法的映射方程

映射方程背后的机制的动态可视化。

扫描二维码下载“湛庐阅读”App，回复书名，观看视频，了解映射方程背后的机制。

最后，L可视为品质函数，对网络的每一种社区划分赋予一个特定的值。为了找到最佳划分，我们必须在所有可能的划分上最小化L。其优化过程的一个广泛使用的实现方式与Louvain算法的步骤I和II类似：把每个节点分配到一个单独的社区，然后系统地合并相邻节点。只要这样的合并能降低L，就能将相邻节点并入模块。每次操作后使用公式**8.45**　更新L。所得的社区再并入超级社区，完成一轮操作，然后算法在这个缩小的新网络上重启。

计算复杂度

Infomap算法的计算复杂度取决于最小化映射方程L的程序。如果我们使用Louvain程序，计算复杂度和Louvain算法一致，即最多O（L

log$L$)或在稀疏图上O (N log$N$)。

总而言之，Louvain算法和Infomap算法提供了实现快速识别社区的工具。它们在基准网络上的准确率与本章讨论的算法相当（图**8-28**）。

8.13
进阶阅读8.D
团渗流的阈值

本节我们推导随机网络上的团渗流的渗流阈值（公式**8.20**），并讨论团渗流算法的主要步骤（图**8-39**）。

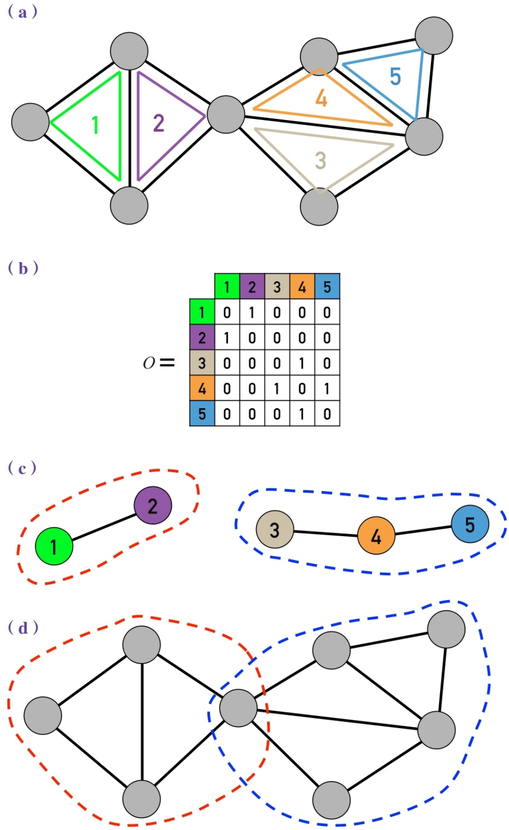


图8-39　团渗流算法

团渗流算法的主要步骤。

（a） 从图中显示的网络开始，我们的目标是识别所有的团。网络中5个k $=3$的团全部被高亮显示。

（b） k $=3$团的重叠矩阵O 。该矩阵被视为一个新网络的邻接矩阵，其节点是原网络的团。该矩阵表明我们有两个连通分支，一个由团（1，2）构成，另一个由团（3，4，5）构成。该网络的连通分支映射为原网络的社区。

（c） 邻接矩阵给出的两个团社区。

（d） 中所示的两个团社区映射到原网络。

当我们以替换一个节点的形式把一个k -团滚到另一个相邻的k -团时，若恰好处于渗流阈值的情况，则能继续滚动的相邻k -团的期望数量应该等于1（图8-20 ）。事实上，该期望值小于1会导致k -团渗流

过早结束，因为从任何k -团开始滚动都会很快停止。因此，簇的大小会呈指数级衰减。另一方面，该期望值大于1会使得团社区无限增大，则系统中一定会出现巨簇（giant cluster）。

上述期望值由下式给出：

$$(k-1)(N-k-1)^{k-1} \quad (8.46)$$

其中$(k-1)$项是可在下一次滚动时被替换的节点个数；$(N-k-1)^{k-1}$项是潜在的滚动目标个数。潜在目标中只有p^{k-1}比例可以真正成为滚动目标，因为滚动带来的新的$k-1$条链接的每一条都必须已经存在，才能构成一个新的团。对于较大的N，公式**8.46** 可简化为：

$$(k-1)Np_c^{k-1}=1$$

即可得公式**8.16** 。

8.14 参考文献

[1] B. Droitcour. Young Incorporated Artists. *Art in America* , 92-97, April 2014.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E.Lefebvre. Fast unfolding of communities in large networks. *J.Stat. Mech* ., 2008.

[3] G.C. Homans. *The Human Groups* . Harcourt, Brace & Co,New York, 1950.

[4] S.A. Rice. The identification of blocs in small political bodies. *Am. Polit. Sci. Rev* ., 21:619-627, 1927.

[5] R.D. Luce and A.D. Perry. A method of matrix analysis of group structure. *Psychometrika* , 14:95-116, 1949.

[6] R.S. Weiss and E. Jacobson. A method for the analysis of

the structure of complex organizations. *Am. Sociol. Rev* ., 20:661-668, 1955.

[7] W.W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res* ., 33:452-473, 1977.

[8] L. Donetti and M.A. Muñoz. Detecting network communities: a new systematic and efficient algorithm. *J. Stat. Mech* ., P10012, 2004.

[9] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *PNAS* , 99:7821-7826, 2002.

[10] L.H. Hartwell, J.J. Hopfield, and A.W. Murray. From molecular to modular cell biology. *Nature* , 402:C47-C52, 1999.

[11] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai,and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science* , 297:1551-1555, 2002.

[12] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *PNAS* , 104:8685-8690, 2007.

[13] J. Menche, A.Sharma, M. Kitsak, S. Ghiassian, M.Vidal, J. Loscalzo, A.-L. Barabási. Oncovering disease-disease relationships through the human interactome. 2014.

[14] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Review Genetics* , 12:56-68, 2011.

[15] G. W. Flake, S. Lawrence, and C.L. Giles. Efficient identification of web communities. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 150-160, 2000.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS* ,

101:2658-2663, 2004.

[17] A.B. Kahng, J. Lienig, I.L. Markov, and J. Hu. *VLSI Physical Design: From Graph Partitioning to Timing Closure* . Springer, 2011.

[18] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal* , 49:291-307, 1970.

[19] G.E. Andrews. *The Theory of Partitions* . Addison-Wesley, Boston, USA, 1976.

[20] L. Lovász. *Combinatorial Problems and Exercises* .North-Holland, Amsterdam, The Netherlands, 1993.

[21] G. Pólya and G. Szeg. *Problems and Theorems in Analysis I* .Springer-Verlag, Berlin, Germany, 1998.

[22] V. H. Moll. Numbers and Functions: From a classical-experimental mathematician's point of view. *American Mathematical Society* , 2012.

[23] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review Letters* , 69:026113, 2004.

[24] M.E.J. Newman. A measure of betweenness centrality based on random walks. *Social Networks* , 27:39-54, 2005.

[25] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol* ., 25:163-177, 2001.

[26] T. Zhou, J.-G. Liu, and B.-H. Wang. Notes on the calculation of node betweenness. *Chinese Physics Letters* , 23:2327-2329, 2006.

[27] E. Ravasz and A.-L. Barabasi. Hierarchical organization in complex networks. *Physical Review Letters* , 67:026112, 2003.

[28] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Physical Review Letters* , 65:066122, 2002.

[29] E. Mones, L. Vicsek, and T. Vicsek. Hierarchy Measure for Complex Networks. *PLOS ONE* , 7:e33799, 2012.

[30] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature* , 453:98-101, 2008.

[31] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics* , P09008, 2005.

[32] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *PNAS* , 104:36-41, 2007.

[33] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review Letters* , 69:066133, 2004.

[34] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *PNAS* , 104:36-41, 2007.

[35] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review Letters* , 70:066111, 2004.

[36] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* , 435:814, 2005.

[37] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E* , 68:065103, 2003.

[38] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech* .,P09008, 2005.

[39] J. Ruan and W. Zhang. Identifying network communities with a high resolution. *Physical Review E,* 77:016104, 2008.

[40] B. H. Good, Y.-A. de Montjoye, and A. Clauset. The performance of modularity maximization in practical contexts. *Physical Review E* , 81:046106, 2010.

[41] R. Guimerá, M.Sales-Pardo, and L.A.N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E* , 70:025101, 2004.

[42] J. Reichardt and S. Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution. *Physical Review E* , 76:015102, 2007.

[43] J. Reichardt and S. Bornholdt. When are networks truly modular? *Physica D* , 224:20-26, 2006.

[44] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS* , 105:1118, 2008.

[45] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *Eur. Phys. J. Special Topics* , 178:13, 2009.

[46] M. Rosvall and C.T. Bergstrom. Mapping change in large networks. *PLOS ONE* , 5:e8694, 2010.

[47] A. Perey. Oksapmin Society and World View. Dissertation for Degree of Doctor of Philosophy. Columbia University, 1973.

[48] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters* , 94:160202, 2005.

[49] J.M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki. A sequential algorithm for fast clique percolation. *Physical Review E* , 78:026109, 2008.

[50] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social

group evolution. *Nature* , 446:664-667, 2007.

[51] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature* , 466:761-764, 2010.

[52] T.S. Evans and R. Lambiotte. Line graphs, link partitions,and overlapping communities. *Physical Review E* , 80:016105, 2009.

[53] M. Chen, K. Kuzmin, and B.K. Szymanski. Community Detection via Maximization of Modularity and Its Variants. *IEEE Trans* . *Computational Social Systems* , 1:46-65, 2014.

[54] I. Farkas, D. Ábel, G. Palla, and T. Vicsek. Weighted network modules. *New J. Phys* ., 9:180, 2007.

[55] S. Lehmann, M. Schwartz, and L.K. Hansen. Biclique communities. *Physical Review E* , 78:016108, 2008.

[56] N. Du, B. Wang, B. Wu, and Y. Wang. Overlapping community detection in bipartite networks. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Los Alamitos, CA,USA: 176-179, 2008.

[57] A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algor* ., 18:116-140, 2001.

[58] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E* , 78:046110, 2008.

[59] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics* , 11:033015, 2009.

[60] S. Fortunato. Community detection in graphs. *Physics*

Reports , 486:75-174, 2010.

[61] D. Hric, R.K. Darst, and S. Fortunato. Community detection in networks: structural clusters versus ground truth. *Physical Review E* , 90:062805, 2014.

[62] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology* , 78:1360-1380, 1973.

[63] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D.Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS* , 104:7332, 2007.

[64] K.-I. Goh, B. Kahng, and D. Kim. Universal Behavior of Load Distribution in Scale-Free Networks. *Physical Review Letters* , 87:278701, 2001.

[65] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, and J.R. Banavar. Universality Classes of Optimal Channel Networks. *Science* , 272:984-986, 1996.

[66] L.C. Freeman. A set of measures of centrality based upon betweenness. *Sociometry* , 40:35-41, 1977.

[67] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *PNAS* , 101:5249-5253, 2004.

[68] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. KDD'07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp. 913-921, 2007.

[69] D.J. Fenn, M.A. Porter, M. McDonald, S. Williams, N.F.Johnson, and N.S. Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007-2008 credit crisis. Chaos, 19:033119, 2009.

[70] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering, in: KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp. 554-560, 2006.

[71] Y. Chi, X. Song, D. Zhou, K. Hino, and B.L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. KDD'07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp. 153-162, 2007.

[72] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B.L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. in: WWW '08: Proceedings of the 17th International Conference on the World Wide Web, ACM,New York, NY, USA, pp. 685-694, 2008.

[73] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth,and evolution. KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp. 44-54, 2006.

[74] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E* , 03G122, 2003.

[75] B. Krishnamurthy and J. Wang. On network-aware clustering of web clients. *SIGCOMM Comput. Commun. Rev* ., 30:97-110, 2000.

[76] K.P. Reddy, M. Kitsuregawa, P. Sreekanth, and S.S.Rao. A graph based approach to extract a neighborhood customer community for collaborative filtering. DNIS '02: Proceedings of the Second International Workshop on Databases in Networked Information Systems, Springer- Verlag, London, UK, pp. 188-200,

2002.

[77] R. Agrawal and H.V. Jagadish. Algorithms for searching massive graphs. *Knowl. Data Eng* ., 6:225-238, 1994.

[78] A.Y. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. KDD’04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY, USA, 2004, pp. 719-724, 2004.

第9章
传播现象

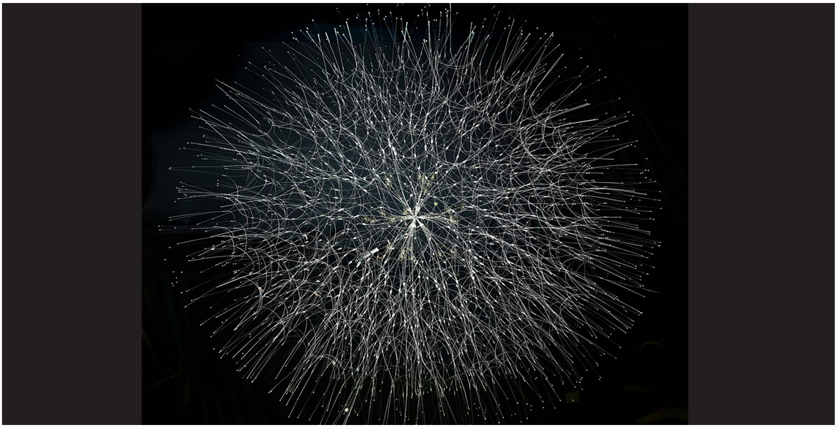


图9-0　比尔·史密斯：传染病传播模型

比尔·史密斯（Bill　Smith）的这幅艺术作品展示了理想情况下的传染病传播模型——一个演化的增长系统。史密斯是伊利诺伊州的一位艺术家。（2009，混合媒体，原作规格213cm×213cm×213cm）

本章索引

9.1 本章概览 / **471**

9.2 流行病建模 / **474**

9.3 网络上的流行病模型 / **481**

9.4 接触网络 / **489**

9.5 其他网络性质的影响 / **494**

9.6 免疫 / **500**

9.7 流行病预测 / **506**

9.8 小结 / **515**

9.9 课后习题 / **517**

9.10 进阶阅读**9.A**
流行病传播的微观模型 / **519**

9.11 进阶阅读**9.B**
SI、**SIS**和**SIR**模型的解析解 / **521**

9.12 进阶阅读**9.C**
针对性免疫 / **526**

9.13 进阶阅读**9.D**
SIR模型和键渗流 / **528**

9.14 参考文献 / **530**

本章致谢

马顿·波什福伊　**MÁRTON PÓSFAI**

萨拉·莫里森　**SARAH MORRISON**

妮科尔·萨迈　**NICOLE SAMAY**

阿迈勒·侯赛尼　**AMAL HUSSEINI**

罗伯塔·西纳特拉　**ROBERTA SINATRA**

菲利普·赫费尔　**PHILIPP HOEVEL**

9.1 本章概览

2003年2月21日晚上，一位医生入住了中国香港京华国际酒店。此前他曾治疗了多名患有“非典”的病人，这种疾病在当时尚无清晰的诊断意见。第二天，在离开酒店后，他前往广东当地一家医院就诊，并在数日后因“非典”去世[1]。

这名医生的入住造成了一系列后果。在他入住当晚，香港京华国际酒店的其他十几名住客也感染了这一疾病——后来被命名为非典型性肺炎（Severe Acute Respiratory Syndrome，简称SARS）。这些携带病毒的客人随后前往河内、新加坡、多伦多等地，引发了这些城市的大规模SARS传染。流行病学家后来跟踪发现，在8 100名确诊患者中，近一半人可以追溯到香港京华国际酒店。而最初携带病毒到香港的那名医生就是一个典型的超级传播者（super-spreader）即在一场疾病传染中导致大规模感染的个体（图**9-1**）。

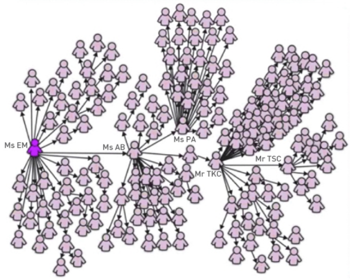


图**9-1**　超级传播者

新加坡确诊的206名非典型性肺炎患者中，有144人曾直接或间接地与5名感染者接触过，这5人中有4人是超级传播者。其中最重要的超级传播者是首例病人——一名医生，他将这一传染病带进了中国香港京华国际酒店。[1]

网络科学研究者通常将超级传播者视为枢纽节点，即在疾病传播所依赖的接触网络中具有大量链接的节点。由于枢纽节点在很多网络中都存在，因此如天花、艾滋病等很多疾病传染记录中都发现了超级

传播者这一角色[2]。在这一章里，我们将引入基于网络的方法来对流行病传染现象进行分析，从而更好地理解和预测这些枢纽节点所产生的真正影响。这一类方法也被称为网络上的流行病模型（network epidemics），这一模型为量化和预测传染性疾病的传播提供了一套有效的分析计算框架。

全球每年因疾病导致死亡的患者中，多达43%的患者的死因是感染了传染性疾病，这些疾病可以通过患者或其分泌物不断传播。药物治疗和疫苗都很难有效地阻止传染性疾病的传播。我们还需要理解病原体是如何在人群中进行传播的，从而反过来决定我们如何管理可用的治疗手段和相应的疫苗。

很多差异巨大的现象都可以被描述为网络中的传播过程。

生物传播

病原体在各个不同的接触网络中的传播是本章讨论的主要内容。例如，流感、SARS、结核等疾病可以通过空气进行传播（两个人在同一个房间中呼吸就可能被传染）；接触性传染病和寄生虫可以通过人们的直接接触进行传播；埃博拉病毒可以通过患者体液进行传播；艾滋病等性传染疾病可以通过性接触进行传播，等等。除此之外，传染性疾病还包括通过致癌病毒（例如人类乳头瘤病毒、人类疱疹病毒第四型）传播的癌症和通过寄生虫（例如臭虫、疟原虫）传播的疾病。

数字传播

计算机病毒是可以进行自我复制并在计算机之间传播的程序。其传播模式同病原体传播类似，但也存在自身的特性——取决于特定计算机病毒背后的具体技术。与此同时，随着智能手机逐渐成为“手持的计算机”，我们也发现了能够在智能手机间传播的手机病毒和蠕虫程序（图**9-2**）。

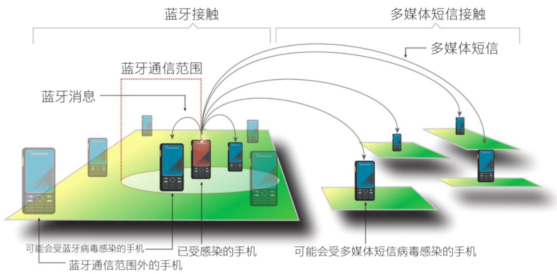


图**9-2**　手机病毒

能够彼此共享程序和数据的智能手机为手机病毒的传播提供了温床。自2004年以来，已经确认的智能手机病毒就有数百种，其短短几年间的发展相当于计算机病毒20多年的发展[3]。手机病毒主要通过两种通信机制进行传播[4]：

蓝牙病毒

受感染的手机将蓝牙病毒传播到蓝牙通信范围内（通常为10~30米）的所有手机。由于物理位置足够近才能用蓝牙建立连接，所以蓝牙病毒的传播能力取决于受感染手机的位置和移动网络的结构。这里的移动网络是指通过人群流动将不同空间位置之间连接起来形成的网络（详见**9.4**节）。因此，蓝牙病毒的传播模式与流感类似。

多媒体短信

受感染的手机通过多媒体短信将病毒传播到通讯录中所有电话号码拥有着的手机上。因为多媒体短信病毒通过社交网络传播，其长程传播模式与受感染手机的物理位置无关。因此，多媒体短信病毒的传播模式与计算机病毒类似。

社交传播

社交网络在传播和接受新概念、知识、商业实践、产品、行为、流言、文化的过程中所起的作用，在社会科学、市场学、经济学等学科中都有广泛的研究[5]，[6]。而Twitter等在线社交平台为我们追踪此类传播现象提供了前所未有的契机。大量的研究者开始聚焦于社交网络中的信息传播问题，探寻诸如一些信息可以传播上百万次而其他信息大多无人问津的原因。

上述所讨论的例子中涉及了不同的传播对象（从病原体到计算机病毒、观点和产品）、不同的传播网络（从计算机网络到社交网络）、不同的时间尺度以及不同的传播机制（表**9-1**）。尽管有诸多不同之处，我们将在本章揭示这些传播过程背后共同遵循的模式和规

律。与此同时，这些传播过程都可以通过基于网络的理论建模框架来进行描述。

表9-1　网络和传播对象

现象	载体	网络
性病	病原体	性行为网络
流言	信息、模因	交流网络
新概念扩散	概念、知识	交流网络
计算机病毒	有害软件、数字病毒	互联网
手机病毒	手机病毒	社交网络、近邻网络
臭虫	寄生虫	酒店与客人构成的网络
疟疾	疟原虫	蚊子与人类构成的网络

病原体、模因或计算机病毒的传播取决于这些载体所处的网络和传播机制。本表列出了一些研究较多的传播现象及其传播载体和所处网络。

9.2
流行病建模

流行病学已经发展出了一套稳定的分析计算框架来进行病原体传播的建模。这一框架主要基于以下两个基本假设：

1.可划分性

流行病模型按照个体所处的发病阶段对其进行分类。最简单的划分方式假设每个个体都处于以下三种阶段之一：

- 易感染(S)：尚未接触病原体的健康个体（图**9-3**）。



图9-3　病原体

“病原体”(Pathogen)一词来源于希腊语“患病，激情”(pathos)和“产生者”(genes)，是指一种有感染能力的载体微生物。病原体是一种致病的微生物，例如病毒、细菌或真菌。图中展示了几种研究较多的病原体，例如人体免疫缺陷（艾滋病）病毒、流感病毒和丙型肝炎病毒。

- 已感染(I)：已接触病原体并能感染他人的个体。
- 已康复(R)：感染后康复的个体，不具备感染能力。

在建模某些疾病的时候还需要考虑一些额外的阶段，例如不能被感染的免疫个体，或者是已接触病原体但尚未感染的潜伏个体。

个体可以在不同阶段之间转化。例如，在流感暴发初期所有人都处于易感染阶段。一旦接触已感染个体，人们就可能转变为已感染阶段。最终，被感染的个体将会康复并获得免疫力，从而不再容易感染这一类型的流感病毒。

2.均匀混合性

均匀混合假设（也称完全混合或近似质量作用）认为每个个体将以相同的概率接触已感染的个体。这一假设使得我们不必知道确切的疾病传播所依赖的接触网络，而是认为任何人都可以感染任何其他人。

这一小节将介绍基于这两个基本假设的流行病建模框架。具体而言，我们将探讨三种常用的流行病模型——SI模型（图**9-4**）、SIS模型（图**9-5**）和SIR模型（图**9-6**），从而对流行病建模的基本组成部分有所了解。

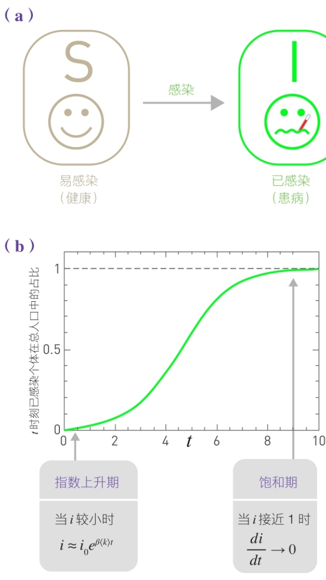


图**9-4**　易感染 - 已感染模型（**SI**模型）

（**a**）　SI模型假设每个个体处于两种状态之一：易感染（健康）和已感染（患病）。模型假设易感染个体接触已感染个体后以概率β 被感染。单向箭头表示一旦一个个体被感染，它将一直处于已感染阶段，不可康复。

（**b**）　根据公式**9.4** 推导出的已感染个体占比随时间变化的过程。早期，已感染个体占比指数上升。最终，所有人都被感染，即$i(\infty)=1$。

易感染 - 已感染模型（**SI**模型）

不妨假设疾病传播的总人口为N。$S(t)$表示t时刻的易感染（健康）个体数，$I(t)$表示t时刻的已感染个体数。在$t=0$时，所有人都是易感染的，即$S(0)=N$；且没有人已经被感染，即$I(0)=0$。假设每个个体在单位时间内有$\langle k\rangle$个可接触对象，且疾病从已感染个体传播到易感染个体的概率为β。我们不禁要问：如果1个个体在$t=0$时被感染，即$I(0)=1$，那么到之后的某一时刻t时将有多少人被感染？

根据均匀混合性假设，已感染个体遇到一个易感染个体的概率是$S(t)/N$，因此该已感染个体将在单位时间内接触$\langle k\rangle S(t)/N$个易感染个体。由于在t时刻共有$I(t)$个已感染个体以概率β传播病原体，因此，在微分时间dt内平均的新增感染个体数$dI(t)$为：

$$\beta\langle k\rangle\frac{S(t)I(t)}{N}dt$$

也就是说，$I(t)$变化率为：

$$\frac{dI(t)}{dt}=\beta\langle k\rangle\frac{S(t)I(t)}{N} \qquad (9.1)$$

在整个章节中，我们都将使用变量：

$$s(t)=S(t)/N,\ \ i(t)=I(t)/N \qquad (9.2)$$

来表示t时刻易感染个体和已感染个体在总人口中的占比。简单起见，我们省略了$s(t)$和$i(t)$中的(t)，将公式**9.1**重新写为如下形式（详见进阶阅读**9.A**）：

$$\frac{di}{dt}=\beta\langle k\rangle si=\beta\langle k\rangle i(1-i) \qquad (9.3)$$

其中$\beta\langle k\rangle$被称为传播率（transmission rate）。为了求解公式**9.3**，我们又将其写为如下形式：

$$\frac{di}{i}+\frac{di}{(1-i)}=\beta\langle k\rangle dt$$

对两侧积分可得：

$$\ln i - \ln(1-i) - \ln C = \beta\langle k\rangle t$$

根据初始条件$i_0 = i(t=0)$可知$C = i_0/(1-i_0)$，从而解得已感染个体在总人口中的占比随时间增长如下：

$$i = \frac{i_0 e^{\beta\langle k\rangle t}}{1 - i_0 + i_0 e^{\beta\langle k\rangle t}} \quad (9.4)$$

公式**9.4** 指出：

- 一开始，已感染个体在总人口中的占比随时间呈指数增长（图**9-4b**）。事实上，已感染个体在早期接触到的都是易感染个体，因此病原体很容易传播。
- 易感染个体占比为原来的1/e（约36%）所需的特征时间（characteristic time）为：

$$\tau = \frac{1}{\beta\langle k\rangle} \quad (9.5)$$

在这里，τ 是病原体在人群中传播速度的倒数。公式**9.5** 指出，提高链接密度$\langle k\rangle$ 或者β 都可以加速病原体传播，缩短特征时间。

- 随着时间推移，已感染个体能接触到的易感染个体越来越少。因此当t 较大时，i 随时间的增长开始放缓（图**9-4b**）。直到所有人都被感染了，疾病传播过程结束。此时$i(t\to\infty)=1$且$s(t\to\infty)=0$。

易感染-已感染-易感染模型（SIS模型）

大多数病原体最终都将被免疫系统或药物治疗消灭。因此我们应该考虑已感染个体康复且不再继续传播的情况。这便是SIS模型。与SI模型相同，SIS模型考虑易感染和已感染两个阶段。不同之处在于，已感染个体以固定概率μ 康复，重新成为易感染个体（图**9-5a**）。SIS模型的动力学方程是公式**9.3** 的扩展：

$$\frac{di}{dt} = \beta\langle k\rangle i(1-i) - \mu i, \quad (9.6)$$

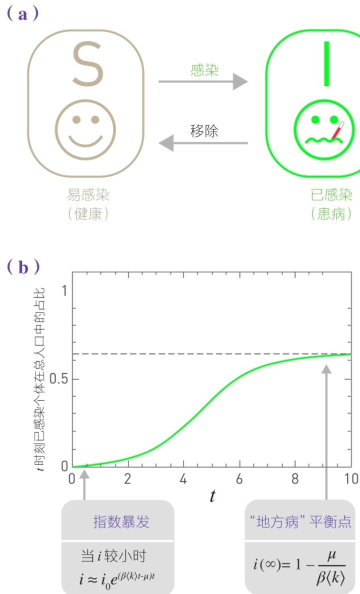


图9-5　易感染 - 已感染 - 易感染模型（**SIS**模型）

（**a**）　与SI模型一样，SIS模型中的个体可以属于两个状态：易感染状态和已感染状态。与SI模型不同的是，SIS模型中的个体可以康复。已感染个体在接受治疗后以速率μ 变成易感染状态。

（**b**）　根据公式**9.7**　推导出的已感染个体占比随时间变化的过程。如果个体可以康复，当t　较大时，系统达到一个所谓的“地方病”平衡点，即已感染个体占比保持恒定，如公式**9.8**　所示的i　(∞)。在该平衡点下，只有部分个体处于已感染状态。当康复速率μ 较高时，已感染个体的数量指数下降并最终趋于零。

其中μ　是康复率，μi　是总人口的康复率。求解公式**9.6**　可得已感染个体在总人口中的占比随时间变化的关系（图**9-5b**　）：

$$i = \left(1 - \frac{\mu}{\beta\langle k\rangle}\right)\frac{Ce^{(\beta\langle k\rangle-\mu)t}}{1+Ce^{(\beta\langle k\rangle-\mu)t}} \quad (9.7)$$

其中，当给定初值条件$i_0 = i(t=0)$时，$C = i_0/(1-i_0-\mu/\beta\langle k\rangle)$。

SI模型认为最终所有人都将被感染。与此不同的是，公式**9.7** 指出在SIS模型中流行病有两种可能的最终形态：

- “地方病”平衡点（$\mu < \beta \langle k \rangle$）

康复概率较低时，已感染个体占比i 随时间的变化服从逻辑曲线（logistic curve），与SI模型类似。不同的是，最终并不是所有人都会被感染，而是常数值$i(\infty) < 1$的人口比例被感染（图**9-5b**）。这意味着在任意时刻，总人口中只有有限的比例受到感染。当达到这样一个“地方病”平衡点时，新增感染个体数等于新增康复个体数。也就是说，已感染个体在总人口中的占比不会随时间变化。令公式**9.6** 中的$di/dt = 0$，可求得：

$$i(\infty) = 1 - \frac{\mu}{\beta \langle k \rangle} \qquad (9.8)$$

- “无病”平衡点（$\mu > \beta \langle k \rangle$）

当康复概率足够高的时候，公式**9.7** 中的指数为负数。因此，i 将随时间指数下降，表明已感染的个体数将会随时间呈指数下降，直至灭绝。这是因为在这种情况下，单位时间内治愈的个体数多于新增的已感染个体。因此一段时间后，这种病原体将从人群中消失。

换言之，SIS模型表明某些病原体将在人口中持续存在，而另一些病原体将很快消亡。为了理解是什么导致了这两种不同的结果，我们推导得出了病原体的特征时间：

$$\tau = \frac{1}{\mu(R_0 - 1)} \qquad (9.9)$$

其中：

$$R_0 = \frac{\beta \langle k \rangle}{\mu} \qquad (9.10)$$

R_0 被称为基本再生数（basic reproductive number）。它表示在所有易感染人群中，被已感染个体所感染的平均个体数。或者说，R_0

是在理想环境下，一个已感染个体所引发的新增感染人数。基本再生数具有很好的参考价值：

- 若R_0超过1，特征时间τ 是正数，传播过程将收敛到“地方病”平衡点。事实上，若每个已感染个体都感染了超过一个健康个体，则病原体可以稳定传播，并在人群中长期存在。R_0越高，传播过程越快。
- 若$R_0<1$，特征时间τ 是负数，传染病将最终消亡。事实上，若每个已感染个体在被治愈前只能感染少于一个健康个体，则病原体无法在人群中长期存在。

因此，基本再生数是流行病学家在面对一个新的病原体时，首先要估算的几个参数之一。它可以被用来估计当前所面临情况的严重程度。表**9-2** 列出了几种已被研究过的病原体的R_0。其中R_0较高的几种病原体的危害较大，例如，每个已感染麻疹的个体可能引发十多人感染。

表**9-2**　基本再生数$\boldsymbol{R_0}$

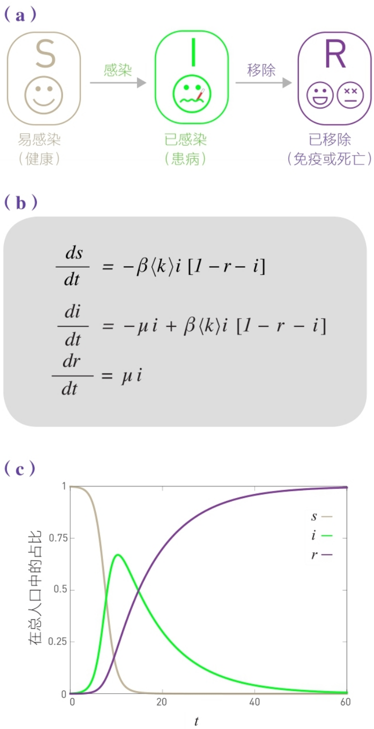
疾病	传播方式	R_0
麻疹	空气	12～18
百日咳	空气飞沫	12～17
白喉	唾液	6～7
天花	社交接触	5～7
小儿麻痹症	粪便–口腔路径	5～7
风疹	空气飞沫	5～7
流行性腮腺炎	空气飞沫	4～7
HIV	性接触	2～5
SARS	空气飞沫	2～5
流感（1918年西班牙流感）	空气飞沫	2～3

再生数（公式**9.10**　）是一个有传播能力的个体在理想情况（所有可接触者都是易感染个体）下所能感染的人数。如果康复个体多于新增感染个体，$R_0<1$的病原体将自然消亡。若$R_0>1$，病原体将扩散传播并在人群中长期存在。R_0越高，传播速度越快。

本表列出了广为人知的几种病原体的R_0。[7]

易感染 - 已感染 - 已康复模型（SIR模型）

对很多病原体（例如大部分的流感病毒）而言，人们感染并康复后会获得免疫能力。这些人群不会再回到易感染阶段，应该从总人口中“移除”。他们不会再被感染，也不会再感染别人，因此不应再计入任何阶段。SIR模型正是用于建模这一动态过程，其性质如图**9-6** 所示。



$$\frac{ds}{dt} = -\beta\langle k\rangle i\,[1 - r - i]$$

$$\frac{di}{dt} = -\mu i + \beta\langle k\rangle i\,[1 - r - i]$$

$$\frac{dr}{dt} = \mu i$$

图**9-6**　易感染 - 已感染 - 已康复模型（**SIR**模型）

（a）　与SIS模型不同，SIR模型中康复的个体处于具有免疫能力的“已康复”状态，不会再成为易感染个体。流感、非典型性肺炎、鼠疫都具有这样的特性，因而需要SIR模型来描述其传播过程。

（b）描述易感染、已感染、已康复个体在总人口中占比（分别表示为s、i、r）随时间演化的微分方程。

（c） 根据图b中方程预测出的s 、i 、r 随时间演化的形态。在SIR模型中，所有人都会经历易感染（健康）、已感染（患病）、已康复（免疫）三个阶段。

综上所述，根据病原体的不同特性，我们需要用不同的模型来刻画流行病传播过程。如图**9-7** 所示，SI、SIS、SIR模型对传染病早期的刻画基本一致：当已感染个体较少时，传染病自由扩散，已感染个体数呈指数级增长。这三种模型的差异主要在于对传染病后期过程的刻画：SI模型中所有个体最终都将被感染；SIS模型中流行病最终将会到达“地方病”平衡点（已感染人口长期处于有限占比），或者逐渐消亡；SIR模型中所有个体最终都会康复。而在这之中，再生数刻画了流行病的最终宿命：$R_0>1$的病原体将在人群中长期存在，$R_0<1$的病原体将自然消亡。

	SI	SIS	SIR
指数时期: 已感染个体数呈指数级增长	$i=\frac{i_0 e^{\beta\langle k\rangle t}}{1-i_0+i_0 e^{\beta\langle k\rangle t}}$	$i=\left(1-\frac{\mu}{\beta\langle k\rangle}\right)\frac{Ce^{(\beta\langle k\rangle-\mu)t}}{1+Ce^{(\beta\langle k\rangle-\mu)t}}$	无闭式解
最终时期: $t\to\infty$ 时达到饱和	$i(\infty)=1$	$i(\infty)=1-\frac{\mu}{\beta\langle k\rangle}$	$i(\infty)=0$
传播阈值: 判断疾病是否能长期存在	无阈值	$R_0=1$	$R_0=1$

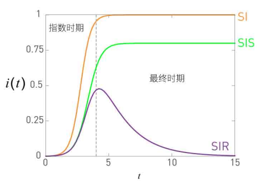


图9-7 对比SI、SIS、SIR模型

SI、SIS、SIR三种模型中已感染个体占比的变化情况。图中可以看到两个不同的时期。

指数时期

三种模型都预测：在传播早期，已感染个体数呈指数增长。对于同样的β ，SI模型预测的传播速度快于SIS模型和SIR模型——SI模型具有最小的τ ，见公式**9.5** 。由于个体可以康复，SIS模型和SIR模型的增长速度要慢于SI模型，因此这两个模型的τ 更大，见公式**9.9** 。当康复速率μ 较高时，SIS模型和SIR模型预测传播过程将到达“无病”平衡点，已感染个体数量随时间指数下降。

最终时期

三种模型给出了不同的长期结果：SI模型认为所有人最终都会被感染，$i(\infty)=1$；SIS模型认为已感染人口将长期处于有限占比，$i(\infty)<1$；SIR模型认为所有已感染个体都将康复，$i(\infty)=0$。

图**9-7** 上方的表总结了每种模型的主要性质。

目前所讨论的模型都是基于均匀混合假设，即任何人都可能感染任何人。事实上，人们只会接触与自己有关的人，即接触网络上的邻居。此外，这些模型都认为每个个体所能感染的人数是$\langle k\rangle$，而忽略了个体在接触网络上的度均有所不同。为了准确地预测流行病传播的动态过程，我们需要考虑接触网络在流行病传播中所起的作用。

9.3 网络上的流行病模型

便利的航空旅行每天让数以百万计的乘客在各大洲之间穿行。而这也加剧了病原体在世界范围的传播。在中世纪，一次流行病需要数年才能蔓延到整个大陆（图**9-8**）；而今天，新的病毒可以在几天之内到达各大洲。因此，我们亟须理解并能预测病原体全球传播的确切模式。

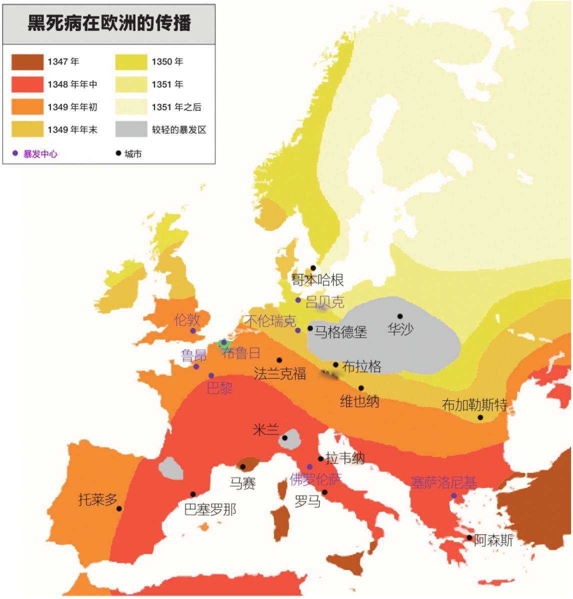


图9-8　大瘟疫

黑死病——人类历史上最具破坏性的大流行病之一，是一场由鼠疫杆菌引起的鼠疫大暴发。图中展示了疾病在数年间逐渐蔓延到整个欧洲的过程。这场流行病始于中亚，1345年冬，蒙古人攻打热那亚领地法卡，将黑死病传入欧洲。之后，病毒可能由商船里常见的黑鼠身上的印鼠客蚤携带，蔓延到地中海和欧洲。它缓慢的传播速度反映了那个时代缓慢的旅行速度。据估计，黑死病杀死了30%~60%的欧洲人口 [8] 。由此产生的破坏引发了一系列宗教、社会和经济动荡，对欧洲历史具有巨大的影响。源自罗杰·曾纳（Roger Zenner）。

上一节讨论的流行病模型均未考虑接触网络（contact network）的结构如何促进了病原体的传播。那些模型假设任何个体都可能接触任何个体（均匀混合假设），且任何个体可接触的人数相差不大，都是$\langle k \rangle$ 。事实上，这两个假设都是错误的：个体只能把病原体传播给能接触到的人，因此病原体是在一个复杂的接触网络上进行传播的；与此同时，接触网络通常是无标度的，因此$\langle k \rangle$ 并不足以充分表征其拓扑结构。

鉴于基础假说是错误的，流行病模型需要进行根本性的修正。这一变化始于罗穆阿尔多·帕斯托尔-萨托拉斯和亚历山德罗·维斯比那尼在2001年的工作：他们扩展了基本的流行病模型，以自洽的方式引

入了底层接触网络的拓扑性质[9]。我们将在本节介绍他们提出的形式化方法，以便了解网络上的流行病模型。

网络上的SI模型

若病原体在网络上传播，链接较多的个体更容易接触到已感染个体，也就更容易被感染。因此模型中必须考虑节点的度，将其作为一个隐变量。得益于度的分块近似理论（degree block approximation），我们可以用节点的度来标识不同的节点，即假设度相同的节点在统计意义下等同（图9-9）。我们用以下公式：

$$i_k = \frac{I_k}{N_k} \tag{9.11}$$



图9-9 度块近似

第9.2节 中讨论的流行病模型把节点按所处状态划分为易感染、已感染、已康复几个部分。考虑到网络拓扑的作用，度块近似增加了一种划分方式，将度相同的所有节点视为一个块。换言之，我们假设度相同的节点有相似的行为。这允许我们对每个度写出一个方程，如公式9.13 。度块近似依然保留基于个体状态的划分：每个个体可能处于易感染（图中空心圆圈）或已感染（图中实心圆圈）阶段，与度无关。

以上公式表示网络上所有N_k 个度为k 的节点中已感染节点的占比。全部已感染节点的占比即为所有已感染的度为k 的节点之和为：

$$i = \sum_k p_k i_k \tag{9.12}$$

对于不同的节点度，我们分别写下对应度k 的SI模型：

$$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k \tag{9.13}$$

这个公式的结构与公式9.3 相同：感染速率正比于β 和度为k 的节点中未感染节点的占比，即$(1-i_k)$。但此外，也存在着一些关键的差异：

- 公式9.3 中的平均度$\langle k\rangle$ 被替换为每个节点真实的度。
- 密度函数Θ_k 表示一个易感染节点k 的邻居中已感染的比例。在均匀混合假说中，Θ_k 即为已感染节点占比i 。然而在网络环境中，一个节点邻居中已感染节点占比取决于该节点的度k 和当前时间t 。
- 公式9.3 用单个方程描述整个系统随时间变化的状态，而公式9.13 联立了$k_{\max}$ 个方程，每个方程对应网络中相应度的节点集合。

出于理论和实践的双重考虑，我们从i_k 的早期行为开始分析。事实上，针对一种新的病原体研发疫苗、特效药和其他医疗手段可能耗时几个月甚至数年。在缺乏治疗手段的时候，唯一可以改变流行病传播进程的措施是尽早采取隔离、限制出行及延缓传染等方法阻止其传播。我们需要准确估计在流行病传播早期已感染个体的数量，以便正确决策采用何种方法、何时开始、在什么程度执行。

在流行病开始传播时i_k 很小，是$\beta i_k k\Theta_k$ 中可以忽略的高阶项。因此公式9.13 可近似为：

$$\frac{di_k}{dt}\approx\beta k\Theta_k \qquad (9.14)$$

正如我们在进阶阅读9.B 中所述，若网络没有度相关性，则Θ_k 函数与k 无关。因此根据公式9.40 ，可将公式9.14 写为：

$$\frac{di_k}{dt}\approx\beta k i_0\frac{\langle k\rangle-1}{\langle k\rangle}e^{t/\tau^{SI}} \qquad (9.15)$$

其中τ^{SI} 是病原体传播的特征时间

$$\tau^{SI}=\frac{\langle k\rangle}{\beta\left(\langle k^2\rangle-\langle k\rangle\right)} \qquad (9.16)$$

对公式9.15 积分，可得度为k 的节点中已感染比例为：

$$i_k = i_0\left(1+\frac{k(\langle k\rangle - 1)}{\langle k^2\rangle - \langle k\rangle}\left(e^{t/\tau^{SI}} - 1\right)\right) \qquad (9.17)$$

公式9.17 给出了如下重要的结论：

- 节点的度越大，越容易被感染。事实上，在任意时间t ，我们可以将公式9.17 写成$i_k = g(t) + kf(t)$。这表明度比较大的节点中有较高的已感染占比（图9-10 ）。

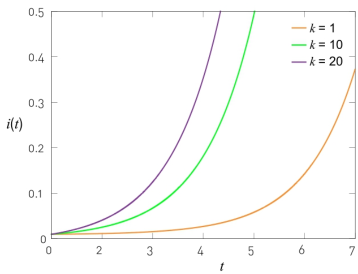


图9-10　SI模型中的已感染节点占比

公式9.17　指出病原体在不同度的节点上以不同速度传播。具体而言，我们写成$i_k = g(t) + kf(t)$，表明在任意时间，大度节点中已感染节点占比高于小度节点中已感染节点占比。图中展示了平均度$\langle k\rangle = 2$的埃尔德什－雷尼网络中度为1、10、20的节点中的已感染节点占比。$t = 3$时，$k = 1$的节点中只有不到3%被感染，$k = 10$的节点中有接近20%被感染，$k = 30$的节点中有接近30%被感染。在任何时刻，几乎所有枢纽节点都被感染，而小度节点极少被感染。总之，病原体在枢纽节点中存活，反过来传播到网络中其余部分。

- 根据公式9.12 ，所有已感染节点占比随时间变化的表达式为：

$$i = \int_0^{k_{max}} i_k p_k dk = i_0\left(1+\frac{\langle k\rangle^2 - \langle k\rangle}{\langle k^2\rangle - \langle k\rangle}\left(e^{t/\tau^{SI}} - 1\right)\right) \qquad (9.18)$$

根据公式9.16 ，特征时间τ 不仅依赖于$\langle k\rangle$ ，还通过$\langle k^2\rangle$ 依赖于网络的度分布。为了完全理解公式9.16 的重要意义，我们对不同的网络分别推导了τ^{SI} 。

- 随机网络

对于随机网络，$\langle k^2\rangle = \langle k\rangle(\langle k\rangle + 1)$，有：

$$\tau_{\mathrm{ER}}^{\mathrm{SI}} = \frac{1}{\beta\langle k\rangle} \qquad (9.19)$$

这等于公式**9.5** 同质网络中的结果。

- $\gamma \geq$ **3**的无标度网络

若流行病传播的接触网络是度指数$\gamma \geq 3$的无标度网络，那么$\langle k\rangle$和$\langle k^2\rangle$都是有限的，从而τ^{SI}也是有限的。此时，流行病传播的动态过程同随机网络上的类似，只是τ^{SI}取值有所不同。

- $\gamma \leq$ **3**的无标度网络

在$\gamma < 3$的情况下，当$N \to \infty$时，$\langle k^2\rangle \to \infty$，此时由公式**9.16** 可得$\tau^{\mathrm{SI}} \to 0$。换言之，病原体在这样的无标度网络上的传播是瞬间暴发的。这大概是网络上的流行病研究最令人意外的结果。

特征时间为零反映了枢纽节点在流行病传播中的重要作用。事实上，如图**9-10** 所示，由于枢纽节点的链接较多从而有较大概率接触到已感染节点，因此它们在无标度网络上是最早被感染的。一旦一个枢纽节点被感染，它就成了一个超级传播者，把病原体传播到网络的其余部分。

- 非同质网络

度异质性的影响并不是无标度网络特有的。事实上，公式**9.16** 指出，只要满足$\langle k^2\rangle > \langle k\rangle(\langle k\rangle + 1)$，就可缩短特征时间$\tau^{\mathrm{SI}}$。也就是说，异质网络可以加速任何病原体的传播。

在SI模型中，随着时间的推移，病原体最终会感染所有个体。因此度异质性只影响了特征时间——表征了病原体扩散到全部人口的速度。为了全面理解网络拓扑结构的影响，我们需要进一步探索网络上

的SIS模型。

SIS模型和传播阈值趋于零

描述网络上的SIS模型的动力学方程是**9.2**节 所述SI模型的直接扩展：

$$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k(t) - \mu i_k \qquad (9.20)$$

公式**9.13** 和公式**9.20** 的不同之处在于康复项$-\mu i_k$ ，这将流行病传播的特征时间变为（见进阶阅读**9.B** ）：

$$\tau^{\text{SIS}} = \frac{\langle k \rangle}{\beta\langle k^2 \rangle - \mu\langle k \rangle} \qquad (9.21)$$

表9-3　网络上的流行病模型

模型	连续介质方程	τ	λ_c
SI	$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k$	$\frac{\langle k \rangle}{\beta(\langle k^2 \rangle - \langle k \rangle)}$	0
SIS	$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k - \mu i_k$	$\frac{\langle k \rangle}{\beta\langle k^2 \rangle - \mu\langle k \rangle}$	$\frac{\langle k \rangle}{\langle k^2 \rangle}$
SIR	$\frac{di_k}{dt} = \beta s_k\Theta_k - \mu i_k$ $s_k = 1 - i_l - r_k$	$\frac{\langle k \rangle}{\beta\langle k^2 \rangle - (\mu + \beta)\langle k \rangle}$	$\frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$

表中列出了任意$\langle k \rangle$ 和$\langle k^2 \rangle$ 网络上的三种基本流行病模型（SI、SIS、SIR）的速率方程，以及对应的特征时间τ 和传播阈值λ_c 。对于SI模型，$\lambda_c = 0$，因为不考虑康复的可能性（$\mu = 0$），病原体会持续传播，直到感染所有易感染个体。τ 和λ_c 的推导过程见进阶阅读**9.B** 。

若μ 足够大，此时特征时间为负，则i_k 随时间呈指数级下降。下降的条件不仅取决于康复概率和$\langle k \rangle$ ，还通过$\langle k^2 \rangle$ 取决于网络异质性。为了说明在什么条件下病原体可以在人群中长期存在，我们对传播率做如下定义：

$$\lambda = \frac{\beta}{\mu} \qquad (9.22)$$

可以看到，传播率仅取决于病原体的生物特征，即传播概率β 和康复概率μ 。λ越大，疾病越容易传播。不过，已感染个体数不会随着λ逐渐增长。只有当传播率超过一定的传播阈值λ_c 时，病原体才能传播。接下来我们分别计算随机网络和无标度网络中的λ_c 。

- 随机网络

如果病原体在随机网络上传播，将$\langle k^2 \rangle = \langle k \rangle (\langle k + 1 \rangle)$ 代入公式**9.21** ，得到病原体在人群中长期存在的条件为：

$$\tau_{\mathrm{ER}}^{\mathrm{SIS}} = \frac{1}{\beta\left(\langle k \rangle + 1\right) - \mu} > 0 \qquad (9.23)$$

结合公式**9.22** 可得：

$$\lambda > \frac{1}{\langle k \rangle + 1} \qquad (9.24)$$

即，随机网络的传播阈值为：

$$\lambda_c = \frac{1}{\langle k \rangle + 1} \qquad (9.25)$$

因为$\langle k \rangle$ 总是有限的，随机网络必有非零的传播阈值（图**9-11** ）。同时，我们还能得到如下关键结论：

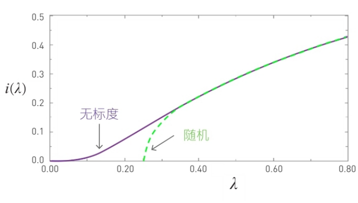


图9-11 传播阈值

SIS模型到达“地方病”平衡点时，已感染个体占比$i(\lambda) = i(t \to \infty)$。绿色曲线表示随机网络，紫色曲线表示无标度网络。随机网络的传播阈值是有限的，表明传播率较小的病原体（$\lambda < \lambda_c$ ）将消亡，即$i(\lambda_c) = 0$。然而，若病原体的传播率超过λ_c ，则该病原体将到达“地方病”平衡点，此后任意时刻存在有限的已感染个体占比。无标度网络中有$\lambda_c = 0$，因此，即使病毒的传播率λ 非常小，它也能在人群中长期存在。

- 若传播率λ超过传播阈值，病原体将持续传播，直至到达“地方病”平衡点。此后任何时间，已感染个体占比均为恒定有限值$i(\lambda)$。
- 若$\lambda < \lambda_c$，该病原体将消亡，即$i(\lambda)=0$。

借助传播阈值我们得以判断一种病原体是否能在人群中长期存在。增大λ会让流行病从寂静状态突变成暴发状态，这也是很多战役中阻止病原体传播的基本认识（**9.6**节）。

- 无标度网络

无论网络有什么样的度分布，我们令公式**9.21** 中的$\tau^{SIS}>0$，可得传播阈值如下：

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \quad (9.26)$$

对于无标度网络，当$N\to\infty$时，$\langle k^2 \rangle$发散，此时网络中传播阈值趋近零（图**9-11** 和图**9-12** ）。这表明即使病毒很难在个体间传播，它们在网络中也将得以传播。这是关于网络上的流行病的第二个基本结论。

流行传播阈值趋近于0，是枢纽节点造成的直接结果。事实上，如果不能在患者康复前感染下一个个体，病原体就会从人群中缓慢消亡（进阶阅读**9.A** ）。在随机网络上，所有节点的度相差不大，均接近$\langle k \rangle$。此时如果传播率低于传播阈值，病原体则无从传播。然而在无标度网络上，即使病原体具有很弱的传染性，一旦枢纽节点被感染，就会将病原体传播到大量节点上，从而使得病原体可以在人群中长期存在。

综上所述，本节的结果表明，考虑网络拓扑大大改变了流行病模型的预测能力。我们得到了以下两个基本结果：

- 在大规模无标度网络上$\tau=0$，这意味着病毒可以立即传播到绝大多数节点。
- 在大规模无标度网络上$\lambda_c=0$，这意味着即使病毒的传播

率很低，也可以在人群中长期存在。

这两个结果都是由枢纽节点可以将病原体传播到大量节点的能力造成的。

注意，这些结果并不仅限于无标度网络。公式**9.16**　和公式**9.26**指出，τ 和λ_c 都依赖于$\langle k^2\rangle$　，因此，上述结果适用于任何有较大的度异质性的网络。换言之，若$\langle k^2\rangle$　大于随机情况下的期望$\langle k\rangle\left(\langle k+1\rangle\right)$，我们将观测到强化的传播过程，即小于传统流行病模型的τ 和λ_c 。这表明病原体的传播将快于传统流行病模型。在设计方案控制流行病传播时，不能忽略这一差异。

本节的结果均基于度块近似理论，即以平均场的方式来处理具体的随时间演化的传染过程。值得注意的是，虽然这个近似简化了表述，但它并不是必须的。我们完全可以用完整的数学形式来求解传染病传播背后的随机过程[11-14]　。这样的完整计算结果表明，由于SIS模型中枢纽节点可以被再次感染，即使在$\gamma>3$的情况下，传播阈值仍会趋近于0。这和我们用平均场方法得出的有限传播阈值显然不同（图**9-12**　）。换句话说，此时枢纽节点的重要程度更甚于我们此前计算的结果。

$2<\gamma<3$	$\lambda_c=0$ $\Theta(\lambda)\sim(k_{\min}\lambda)^{(\gamma-2)/(3-\gamma)}$ $i(\lambda)\sim\lambda^{1/(3-\gamma)}$
$\gamma=3$	$\lambda_c=0$ $\Theta(\lambda)\approx\dfrac{e^{-1/k_{\min}\lambda}}{\lambda k_{\min}}(1-e^{-1/k_{\min}\lambda})^{-1}$ $i(\lambda)\sim 2e^{-1/k_{\min}\lambda}$
$3<\gamma<4$	$\lambda_c>0$ $i(\lambda)\sim\left(\lambda-\dfrac{\gamma-3}{k_{\min}(\gamma-2)}\right)^{1/(\gamma-3)}$
$\gamma>4$	$\lambda_c>0$ $i(\lambda)\sim\lambda-\dfrac{\gamma-3}{k_{\min}(\gamma-2)}$

图**9-12**　**SIS**模型的渐近性

到达“地方病”平衡点时，已感染个体占比$i(\lambda)=i(t\to\infty)$取决于底层网络结构和疾

病特征β和μ。图中总结了度指数为γ的无标度网络上的传播阈值λ_c、密度函数Θ(λ)和i(λ)的关键性质。结果表明，只有在γ＞4时，无标度网络上的流行病传播才会收敛到传统流行病模型的结果。[10]

9.4 接触网络

网络上的流行病学研究表明，病原体传播的速度取决于相应的接触网络的度分布。事实上，我们发现特征时间τ和传播阈值λ_c都受到$\langle k^2\rangle$的影响。若病原体在随机网络上传播，则上述结论都不成立。此时，网络上的流行病传播与**9.2**节讨论的传统的流行病模型一致。本节我们将介绍流行病中的几种接触网络的结构，为证明网络度异质性的显著存在提供直接的经验性证据。

性传播疾病

引起艾滋病的人类免疫缺陷病毒主要通过性接触传播，因此，其接触网络刻画了谁与谁存在性关系。一项针对瑞典人的性习惯调研最早揭示了性关系网络的结构[15]。通过访谈和问卷调查，研究人员收集了4 781名随机选择的18~74岁的瑞典人的信息。调查参与者只需要估计出自己一生中所有性伴侣的个数，而不必说出他们的名字。通过这个数据，研究人员得以重建性关系网络的度分布[16]，并发现幂律可以很好地近似表达这一度分布（图**9-13**）。这首次为无标度网络与病原体传播的相关性提供了经验证据。这一发现在之后英国、美国和南非的数据中也得到了验证[17]。

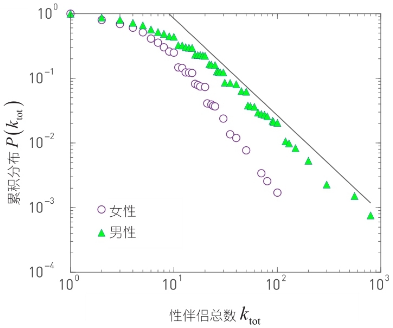


图9-13 性关系网络

此图描述的是，性伴侣总数的累积分布，来自1996年瑞典的一份访谈报告 [15] 。女性的性伴侣总数分布尾部(k >20)服从γ =3.1±0.3的幂律分布，男性则在20<k <400的区间上服从γ =2.6±0.3的幂律分布。注意，男性的平均性伴侣个数高于女性，这可能是社会偏见所致：男性倾向于夸大性伴侣的数量，女性倾向于隐瞒性伴侣的数量。[16]

性关系网络的无标度属性表明，大部分个体的性伴侣数都较少，而少数个体则在一生中拥有数百个性伴侣。因此，性关系网络具有较大的$\langle k^2\rangle$ ，从而τ 和λ_c 较小。

边栏9.1

性枢纽节点

据说，性枢纽节点是真实存在的。例如，20世纪80年代名人堂篮球运动员张伯伦（Wilt Chamberlain）惊人地声称曾与2万名女性发生性关系。“是的，这是真的，2万名不同的女性，”他在自传中写道[18] ，“在我现在的年龄，这等于是从15岁开始每天与1.2名女性发生性关系。”在艾滋病的历史上，航空乘务员吉坦·杜加斯（Geetan Dugas）的故事是有据可查的 [19] ，他有大约250名同性伴侣。他被称为“零号病人”，因为频繁旅行传播艾滋病而成为男同性恋社区的超级传播者。甚至在高中的恋爱关系网络中，也存在枢纽节点（图**9-14** ）。

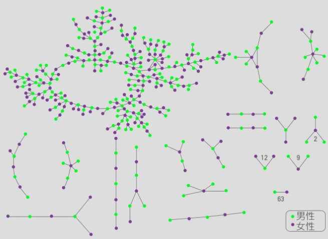


图9-14　一所高中的恋爱关系链

美国中西部高中学生的恋爱和性行为链。每个圆圈表示一名学生，每条链接表示其在接受访问前6个月里有恋爱关系。数字表示同样结构的子图的个数：有63对情侣与这个网络的其他部分不相连。[20]

空气传播疾病

流感、非典型性肺炎、H1N1甲型流感等空气传播疾病的接触网络刻画了人们在物理空间的近邻性。我们从两个层面讨论这一接触网络的结构。首先，全球旅行网络作为几个大规模流行病预测工具的输入，让我们得以预测病原体如何在世界范围传播（**9.7**节）。其次，我们可以用数字徽章探测到接触网络的局部特性，即与一个人直接接触的个体数目。

全球旅行网络

若想预测病原体的传播，我们需要知道已感染个体旅行了多远。手机的使用提供了个体移动性的直接信息，极大地促进了我们对于个体旅行模式的理解[21-24]。在流行病中，航空旅行的移动性数据被研究得最多，因为它决定了病原体全球传播的速度。因此，表示机场之间直飞航线的航空运输网络，在建模和预测病原体传播方面有着关键作用[25-27]。如图**9-15** 所示，该网络是度指数$\gamma=1.8$的无标度网络。度指数较低是因为该网络不是简单网络：两个机场之间可以存在多个航班。链接权重的分布也是类似的幂律分布，这意味着在大多数机场之间旅行的乘客数都较少，但在某些机场之间旅行的乘客非常多。正如我们在**9.5**节 的讨论，这一异质性对病原体传播起着关键作用。

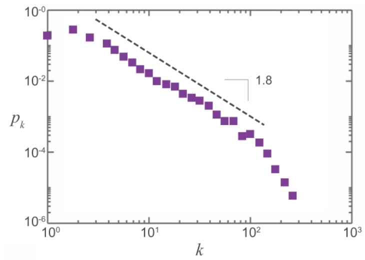
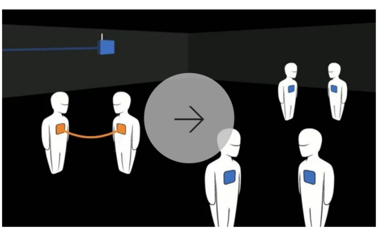


图9-15　航空运输网络

航空运输网络的度分布可以用$\gamma=1.8\pm0.2$的幂律分布很好地近似。数据来自国际航空运输协会的数据库，包含2002年全世界的机场列表和它们之间的直飞航线。这一带权网络中的节点是$N=3\ 100$个最大的机场，还有$L=17\ 182$条表示直飞航线的链接，涵盖了全世界99%的航空运输。[25]

局部接触模式

很多通过空气传播的疾病都需要人们面对面地交流，从而进行传染[28-31]。这些交流接触情况可以用射频识别装置（RFID）[29]，[31]，以及基于手机的社交徽章[32]，[33]或其他无线技术[34]来监测。射频识别装置是一种数字徽章，可以检测佩戴通胀徽章的个体之间的邻近程度（在线资源9.1）。这一技术已被应用于多种场景，例如可以检测科学馆内14 000名访客在3个月内的交流接触情况或检测100名参会人员在为期三天的学术会议上的交流接触情况[29]。图9-16展示了一个利用射频识别装置检测到的网接触络——捕获了高中学生和老师在两天里的交流情况，有如下几项重要的发现：



在线资源9.1　利用射频识别技术检测网络

介绍射频识别技术及其应用于识别人际交互的视频。

扫描二维码下载“湛庐阅读”App，回复书名，观看视频，了解射频识别技术。

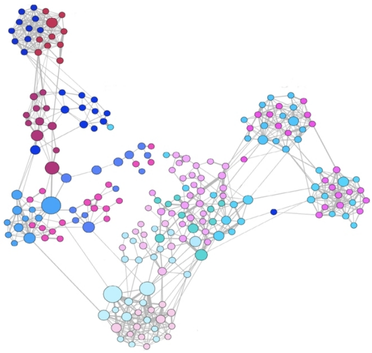


图9-16　面对面交流

此图描述的是，利用射频识别标签绘制出的面对面交流网络，包含一个学校10个班里的232名学生和10名教师的交流 [31] 。采集射频识别信号的场合决定了利用标签获得的结构。例如，图中所示的学校网络展示出清晰的社区。与之相反，博物馆参观人群之间交流的网络几乎是线性网络。再者，小型会议的参会人员之间交流的网络非常稠密，因为大多数人都会和大多数人交谈 [29] ， [31] 。

- 射频识别标签只能检测佩戴同种徽章且面对面的个体之间的交流情况，限制了可以检测到的接触者数量。因此，这些研究中绘制出的接触网络通常服从指数形式的度分布。
- 每次面对面交流的时长服从跨越好几个量级的幂律分布。也就是说，大多数交流都很简短，但也存在少部分长时间的交流。这一结果表现出了时间上的阵发模式 [35] ，而这对病原体传播也有关键影响（**9.5**节 ）。
- 链接权重表示两个个体交流的总时长，同样服从幂律分布。这说明人们把大部分时间都花在了与少数几个人的交流上，这同样对传播模式有重要影响（**9.5**节 ）。

- 对于大部分空气传播的疾病来说，其实只要在空间上接近即可传播，不需要彼此面对面。例如，某个个体在电梯里站在了一名易感染者背后，这已经足以使得该个体感染非典型性肺炎或H1N1甲型流感，但却并不会被射频识别设备记录。

总而言之，射频识别设备提供了详细的局部接触者的时空信息。为了使研究更有效，研究人员同时还应该扩大范围，获得更多种类的数据，例如使用基于移动手机的检测[36]。

地理位置网络

对于很多空气传播的疾病来说，其接触网络被称为地理位置网络。地理位置网络的节点是位置，两个节点之间的链接表示常常在两个位置之间移动的人口。基于代理人的模拟过程可知，地理位置网络是厚尾的[37]：商场、机场、学校或超市扮演了枢纽节点的角色，连接了大量的家庭、办公室等人流较少的地理位置。因此，一旦病原体感染了某个枢纽节点，疾病很快将大量地传染到其他节点。

数字病毒

针对感染计算机和智能手机的数字病毒的研究，是流行病现象中越来越重要的一项应用。接下来我们要讨论其对应的接触网络，它取决于相应的数字病原体的传播模式。

计算机病毒

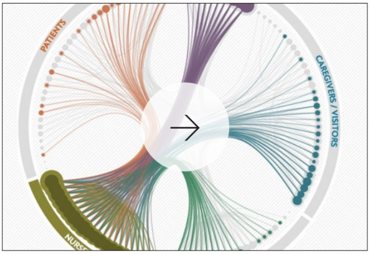
计算机病毒展现出了和生物病毒一样的多样性：不同的病毒及传播机制，其对应的接触网络可以有很大的差异。很多计算机病毒通过电子邮件的附件进行传播。当一个用户打开附件，病毒就会感染该用户的计算机，并将自身的副本发送到在这台计算机上找到的所有电子邮件地址。因此，这类病毒的接触网络是电子邮件网络。正如我们在**表3-1** 中所述，电子邮件网络是无标度网络[58]。另一些计算机病毒

则利用各种通信工具在网络上传播，其传播方式反映了互联网的互联模式，也是无标度的（表**3-1**　）。此外，还有一些恶意软件会通过扫描IP地址在全连通的网络上传播病毒。

手机病毒

手机病毒通过多媒体短信和蓝牙传播（图**9-2**　）。一个多媒体短信病毒会将自身的副本发送给被感染手机通讯录中的所有电话号码。因此，多媒体短信病毒利用的是移动通信背后的社交网络。如表**3-1**所示，手机通话网络是度指数较大的无标度网络。手机病毒也可以通过蓝牙传播，将自身的副本发送给空间位置接近的蓝牙范围内的所有易感染手机。如前所述，这个共同位置网络也是具有高异质性的[4]。

总而言之，过去10年的技术进步使我们可以探索生物病毒和数字病毒传播的各种接触网络的结构，包括性接触网络和基于近邻性的接触网络等（在线资源**9.2**　）。其中包括电子邮件网络、互联网、性关系网络在内的大部分网络都是无标度的。其余的网络，例如共同位置网络，虽然它们的度分布并不能用简单的幂律拟合，但依然表现出了显著的度异质性——$\langle k^2 \rangle$ 较高。这表明此前章节的分析结果适用于病原体在大多数接触网络上的传播。也就是说，底层的异质接触网络甚至可以让传播能力较弱的病毒在人群中轻易地进行传播。



在线资源**9.2**　医院里的暴发

细菌对抗生素的耐药性是全球卫生系统面临的一个严峻挑战。耐药细菌在医院和医疗保健机构尤其常见。《科学美国人》的交互视频描述了细菌在医院中的暴发。

扫描二维码下载“湛庐阅读”App，回复书名，了解细菌在医院暴发的详情。

9.5 其他网络性质的影响

到目前为止我们的模型都是简单的：假设病原体在无权网络上传播，且这一网络完全由度分布决定。然而，真实网络还有一些性质无法通过p_k 刻画，例如度相关性或社区结构。此外，网络的链接通常是有权重的，交互行为也存在有限的时长。在这一节中，我们将探讨这些性质对病原体传播的影响。

含时网络

我们通过社交链接所进行的交流大多是短暂且不频繁的。由于病原体只能在实际接触发生时传播，我们还需要一个更精确的模型框架来考虑每次交互的发生时刻和时长。忽略交互的时间性可能导致具有误导性的结论[39-41]。例如，我们通过图**9-17a** 中的个体交互行为聚合得到图**9-17b** 的静态网络。在这一聚合网络上，病原体从D传播到A的机会与从A传播到D的机会相同。然而，当我们进一步考虑交互时间，可以发现，虽然从A开始的感染可以传播到D，但从D开始的感染并不能传播到A。因此，要准确地预测传播过程，必须考虑病原体是在一个含时网络上传播的事实，这也是网络科学领域越来越活跃的一个研究课题[40-43]。如果忽略这些接触模式的含时性，我们通常会高估流行病暴发的速度和程度[42]，[43]。

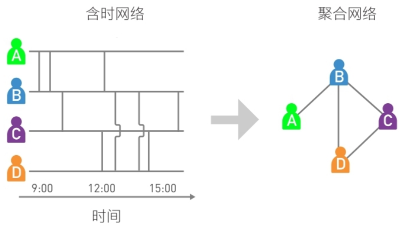


图9-17　含时网络

网络上的大多数交互都不是永远持续的，而只会持续有限的时长。因此我们必须将底层网络视为含时网络。含时网络是网络科学领域越来越活跃的一个研究课题。

（a）含时网络

4个个体交互的时间轴。每一条竖线表示2个个体相互接触。如果首先被感染的人是A，病原体可以从A传播到B，然后到C，最后到达D。然而，如果首先被感染的人是D，疾病会传染C和B，但不会到达A。这就是从A到D有一条含时路径的原因。

（b）聚合网络

此图是通过合并图a中的含时交互行为得到的网络。如果我们只有聚合网络的表达，病原体从任意节点出发都可以到达所有个体。[40]

阵发式接触模式

在9.2节　和9.3节　讨论的理论方法中，我们假设两个连通节点之间的交互时间是随机的。这表明相邻两次交互的间隔时间服从指数分布，交互行为随机而均匀地分布在时间轴上（图9-18a　~c　）。而实践测量发现并非如此，在大多数社交系统中间隔时间服从幂律分布[35]，[44]（图9-18d　~f　）。这表明两个个体之间的接触行为中具有典型的频繁交互时段，即大量接触行为在相对较短的时长里出现。与此同时，幂律分布也意味着会出现很长的时间间隔。因此接触模式在时间轴上具有非均匀的阵发特性（图9-18d、e　）。

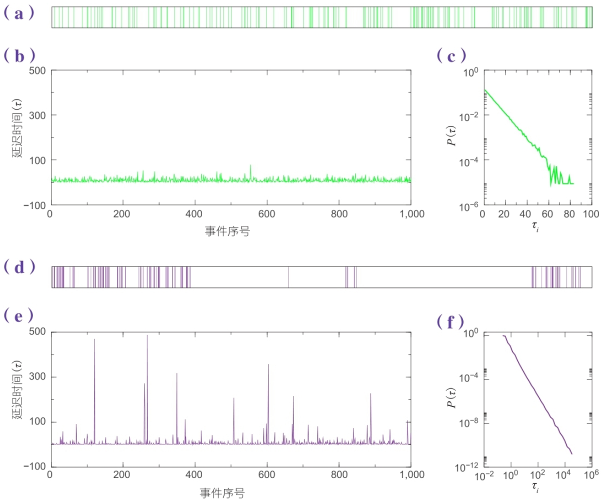


图9-18　阵发式的交互

（a）　如果个体的活动模式是随机的，间隔时间应服从泊松过程（Poisson process），即假设事件在任意时刻以均等概率q　发生。横轴表示时间，每一条竖线表示随机时刻发生的一次事件。观测到的各个时间间隔的长短相差不大，长的时间间隔很罕见。

（b）1 000次连续随机事件的时间间隔为τ_i。可以看到长的时间间隔是很短的。竖线的高度表示图a中的时间间隔长短。

（c）在一段固定时间内观测到恰好n 次事件的概率服从泊松分布$P(n,q)=e^{-qt}(qt)^n/n!$，表明时间间隔分布服从$P(\tau_i)\sim e^{-q\tau_i}$。如线性－对数坐标下的图所示。

（d）　此图描述的是，含时模式中的事件序列。其时间间隔服从幂律分布，虽然大部分事件之间的间隔极短（阵发行为），亦存在几次超长时间间隔，对应于接触模式中的长时间沉默。这一时间序列并不像图a一样是均匀的，而是有阵发性的。

（e）1 000次连续事件的等待时间τ_i。平均时间间隔被设置为图b中泊松过程的平均时间间隔。较大的尖峰对应超长的延迟时间。

（f）此图描述的是，对于图d和图e中的阵发过程，延迟时间分布$P(\tau_i)\sim\tau_i^{-2}$。[35]

阵发性交互在很多与传播现象相关的接触过程中都被发现，从电子邮件交流到通话、性接触等。一旦阵发出现，动态传播过程就会被改变[43]。具体而言，服从幂律分布的间隔时间增大了特征时间τ，因此，受感染个体的数量的衰减速度将慢于随机接触模式下所预测的

结果。例如，若连续两封电子邮件之间的时间间隔服从泊松分布，电子邮件病毒的衰减服从$i(t) \sim \exp(-t/\tau)$，衰减时间为$\tau \approx 1$天。然而，真实情况下的衰减会慢很多，衰减时间为$\tau \approx 21$天，这与我们使用幂律间隔分布理论的预测相吻合[43]。

度相关性

如第**6**章 所讨论的，许多社交网络是同配网络，大度节点倾向于和其他大度节点相连。那么这样的度相关性是否会影响病原体的传播呢？计算表明，度相关性并不能改变网络上的流行病传播的关键性质，但会改变病原体在网络上传播的速度：

- 度相关性改变了传播阈值λ_c：同配相关降低λ_c，而异配相关增加λ_c[45]，[46]。
- 尽管λ_c会改变，但无论无标度网络是同配、中性还是异配的，由于其二阶矩的发散性，SIS模型的流行度传播阈值在无标度网络上仍会趋近于零[47]。因此**9.3**节 的基本结论不受度相关性影响。
- 若枢纽节点首先被感染，则同配网络会加速病原体的传播，而异配网络会减缓传播过程。
- 最后，研究发现在SIR模型中，同配相关性会降低流行病的流行程度，但会增加其平均的暴发时长[48]。

链接权重和社区

在本章中我们假设所有链接都是等权的，重点关注病原体在无权网络上的传播。但真实网络中的链接权重可能有很大的不同，这一异质性对于传播现象有重要的影响。事实上，个体与另一个已感染个体接触次数越多，被感染的概率越高。

同样，以前我们忽略了病原体传播网络中的社区结构。然而，社区的存在（第**8**章 ）使得社区内的节点会进行重复交互，从而改变动态传播过程。

手机网络恰好为我们提供了探寻链接权重以及社区在传播中所起作用的机会[49]。我们假设在$t=0$时随机挑选一个个体，并给予其一些信息（感染）。在接下来的每一个时刻，这名已“感染”个体i以概率$p_{ij}\sim\beta w_{ij}$把信息传递给他的联系人j，其中β是传播概率，w_{ij}是链接权重——定义为i和j通话的分钟数。事实上，两个个体的通话越多，他们传递信息的概率就越大。为了理解链接权重在传播过程中所起的作用，我们同时也考虑控制网络上的传播过程，即连接结构相同但链接权重被设置为等权$w=\langle w_{ij}\rangle$的网络。

如图**9-19a**所示，控制网络上的信息传播明显更快。真实网络中观测到的较慢速度表明，信息被困在了社区内部。事实上，如第**8**章所讨论的，强链接倾向于出现在社区内部，而弱链接倾向于出现在社区之间[50]。因此，在强链接作用下，一旦信息到达社区的某一个成员，它将迅速传播到这个社区里所有其他成员。然而，因为社区之间的链接是弱链接，信息很难离开当前的社区。因此，当信息快速感染整个社区之后，会有很长的一段时间被困在该社区内部。而在所有链接都是等权的控制网络中，社区之间的联通增强了，因而信息不会再被困在社区内部。

图**9-19b**、**c**展示了手机通话网络中小范围内的一种传播情况，以表现真实网络和控制网络中的传播过程的差异。在控制网络的模拟中，信息倾向于沿着最短路径传播。当考虑链接权重时，信息沿着较长的链接权重较大的骨干网络传播。例如，图**9-19b**中信息极少到达网络的下半部分，而这部分在图**9-19c**的控制网络模拟中总是能被传染到。

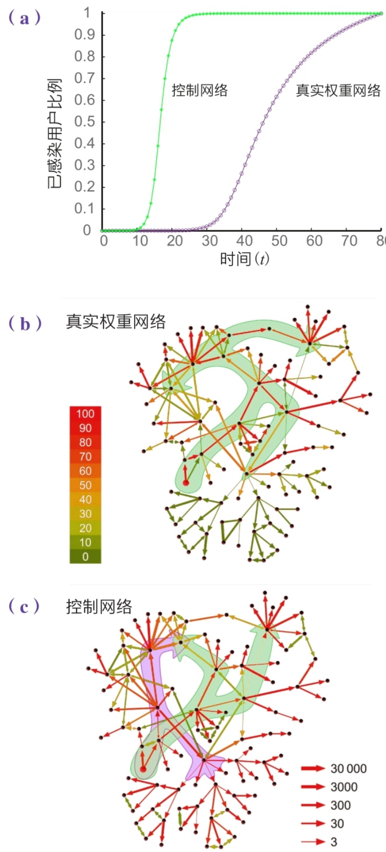


图9-19　手机网络上的信息传播

带权手机通话网络上的信息传播。节点传递信息给邻居节点的概率正比于它们之间的链接权重——彼此之间以分钟计算的通话时长。

（a）　已感染节点占比随时间变化的函数。蓝色圆圈表示真实权重网络上的传播，绿色星号对应控制网络——所有链接的权重相等。

（b）　真实权重网络中一个小规模邻域上的传播。红色节点发出信息，箭头的权表示链接权重。箭头的大小正比于1　000次模拟中尝试传播的次数，颜色表示成功传播的次数。阴影轮廓凸显了真实权重网络和控制网络上传播方向的不同。

（c）与图b相同的控制网络，所有链接的权重相等，$w = \langle w_{ij} \rangle$。[49]

复杂传染

社区在传播中起到了多重的影响，从引发全球级联效应[51]，[52]到改变个体活动[53]。

思想、行为等个体间的文化传播进一步凸显了社区的重要作用[54]。文化传播吸引了包括市场营销[5]，[55]、网络科学[56]，[57]、通信[58]、社交媒体[59-61]等领域的大量关注。病原体和文化在传染上具有不同的模式，这也促使我们对简单传染和复杂传染进行系统性的区分[54]，[62]，[63]。

迄今为止探讨的传染过程都是简单传染：已感染个体和易感染个体接触即可进行病原体的传染。但文化、产品和行为的传播往往被描述为复杂传染——很多人并不会在第一次接触一种新的文化、产品或行为模式的时候就接受它们。相反，接受新事物需要不断的强化[64]，即与已感染个体反复接触。例如，一个人的朋友中使用手机的比例越高，他越可能买手机。

在简单传染的场景中，社区会困住信息和病原体，减缓传播速度（图**9-19a**）。这一效应在复杂传染的场景中正好相反：社区里的冗余链接提供了社交的强化，使得个体暴露在多个已感染个体面前。因此，社区反而可以孵化文化、商品或行为模式，使其更容易被接受。

简单传染和复杂传染的区别可以用Twitter数据很好地说明。推文或短消息通常带有关键字标签，这些标签可视为一种文化。Twitter上的用户可以通过关注其他用户来接收他们发布的推文，也可以把推文转发给自己的关注者，或在推文中提及他人。根据测量，大多数标签都被困在特定的社区内，这是复杂传染的典型标志[54]。特定社区内对一个标签的频繁提及正是社交强化的证据。相反，病毒式标签会在多个社区之间传播，其传播模式类似于生物病原体。通常来说，标签到达的社区越多，证明其越具有感染性（图**9-20**）。

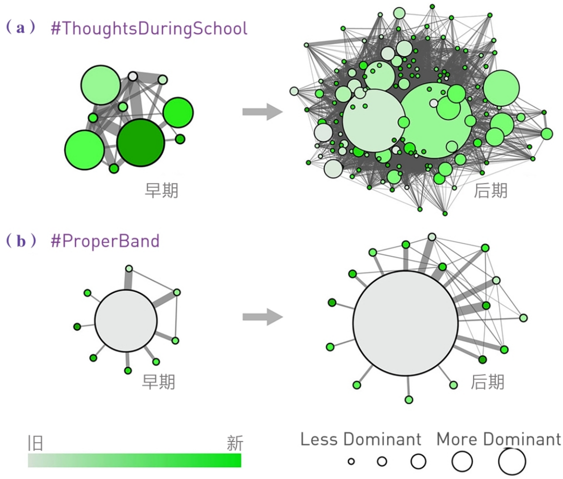


图9-20　简单传染和复杂传染

此图描述的是，Twitter关注者网络的社区结构。每个圆圈代表一个社区，圆圈大小正比于社区发送推文的数量。社区颜色代表所关注的主题标签（模因）最早在这个社区内出现的时间：首先使用该主题标签的社区是浅色，较晚使用的社区是深色。

（**a**）简单传染

主题标签#ThoughtsDuringSchool的早期（30条推文，左图）和晚期（200条推文，右图）传播情况。图中展示了病毒式模因随时间演化的过程。模因很容易从一个社区跳到另一个社区，感染了大部分社区，服从生物病原体的传染模式。

（**b**）复杂传染

主题标签#ProperBand的早期（左图）和晚期（65条推文，右图）传播情况。图中展示了非病毒式模因随时间演化的过程。推文被困在少数几个社区内，很难脱离出去。这是强化的典型标志，表明这个模因属于复杂传染。[54]

综上所述，其他一些网络性质，如度相关性、链接权重、阵发性接触模式等，都会影响病原体在网络上的传播。如我们在本节所讨论的，一些网络性质减慢了病原体的传播，另一些则促进了其传播。因此，当我们预测真实病原体传播时，必须考虑这些网络的性质。此外，这些性质不仅会影响典型的传染性疾病的传播，还会对肥胖等非传染性疾病的传播造成影响（边栏**9.2**）。

边栏9.2

我们的朋友让我们发胖？

流感、非典型性肺炎、艾滋病等传染性疾病通过病原体传播来扩散。但社交网络是否也会帮助非传染性疾病传播？确实如此，最近有证据表明，社交网络会影响肥胖、幸福感和行为模式（例如戒烟）的传播 [65] ， [66] 。

肥胖可以通过个体的体重指数（BMI）来判断，取决于基因、饮食、锻炼等多种因素。有证据表明，朋友也是一个重要因素。一项关于5 209名男性和女性的社交网络的分析发现，如果我们的朋友中有一个肥胖的人，我们在未来2~4年内超重的风险会增加57% [65] 。如果我们最好的朋友超重，那么这一风险会增加3倍：我们超重的概率升高了171%（图**9-21** ）。从实践角度来看，肥胖可被视为与流感、艾滋病同样的接触传染性疾病，虽然并不存在“肥胖病原体”。



在线资源**9.3**　社交网络上的传播

“如果你的朋友肥胖，你肥胖的风险比普通人高45%……如果你朋友的朋友肥胖，你肥胖的风险会提高25%……如果你朋友的朋友的朋友肥胖（你可能都不认识他），你肥胖的风险会提高10%。只有当这个肥胖的人是你朋友的朋友的朋友的朋友时，他的体重才和你没关系。”在视频中，尼古拉斯·克里斯塔基斯（Nicholas Christakis）解释了健康模式是如何通过社交网络传播的。

扫描二维码下载“湛庐阅读”App，回复书名，进一步了解社交网络上的传播模式。

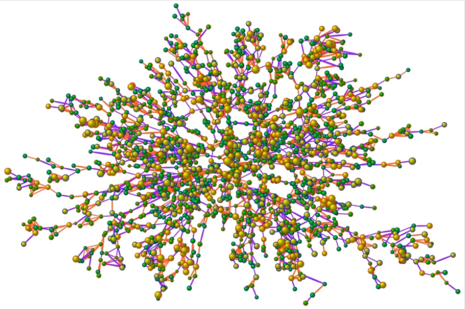


图9-21　肥胖网络

此图是一个社交网络的巨连通分支。这个社交网络展示了2 200名参加弗雷明汉心脏病研究（Framingham Heart Study）项目的个体之间的朋友关系。每个节点代表一个个体：蓝色表示男性，红色表示女性。节点大小与个体的体重指数成正比，黄色节点表示肥胖个体（体重指数≥30）。紫色链接表示朋友或夫妻关系，橙色链接表示家庭关系（例如兄弟姐妹）。图中可以看到肥胖的个体和不肥胖的个体分别形成的簇。分析指出，同配性（体形相似的个体倾向于彼此成为朋友）无法解释这些簇。研究人员将其归因于复杂传染过程——刻画肥胖沿着社交网络的链接进行“传播”的过程。[65]

9.6 免疫

免疫策略用于指导如何在人群中进行疫苗、治疗手段和药物的分配。理想状态下，只要有治疗手段或疫苗，就应该让每个已感染个体接受治疗，并让每个有接触病原体风险的个体接种疫苗。然而，由于难以覆盖全部有感染风险的个体（通常出于成本考虑），以及治疗手段存在副作用（真实的或心理作用的），免疫策略无法做到完全覆盖所有人群。基于这些限制，免疫策略的目标是尽可能有效地分配可用的疫苗和治疗手段，以尽可能降低流行病的威胁。

免疫策略会受传统流行病模型的预测结果的指导：如果一种病原体的传播率λ低于传播阈值λ_c，这种病毒会自然消亡（图**9-11**）。然而，在无标度网络中，这个传播阈值会趋向零，使得我们对这一原则的有效性暂且存疑。事实上，当传播阈值为零时，免疫策略无法将λ降到λ_c以下。本节将讨论如何通过我们对网络拓扑的理解，来设计

有效的、基于网络的免疫策略，从而抵抗传播阈值为零的影响。

随机免疫

免疫的主要目标是保护个体免受感染。而它的另一目标与之同等重要，即降低病原体在人群中传播的速度。为了说明这一效果，让我们考虑从总人口中随机选择g比例的个体免疫的情况[8]。

假设病原体传播遵循SIS模型（公式**9.3**）。接受免疫的节点对病原体是不可见的，只有余下的$(1-g)$比例的节点可能接触并传播这一疾病。因此，每个易感染节点的有效度从$\langle k\rangle$变成了$\langle k\rangle(1-g)$，使得病原体的传播率从$\lambda=\beta/\mu$降到了$\lambda'=\lambda(1-g)$。接下来让我们看看这一改变在随机接触网络和无标度接触网络上的影响。

- 随机网络

如果病原体在随机网络上传播，此时较高的g可以使得传播率λ'低于传播阈值（公式**9.25**）。可以达到这一效果所需的最低免疫率g_c满足

$$\frac{(1-g_c)\beta}{\mu}=\frac{1}{\langle k\rangle+1}$$

可得

$$g_c=1-\frac{\mu}{\beta}\cdot\frac{1}{\langle k\rangle+1} \qquad (9.27)$$

因此，如果能够通过接种疫苗使免疫个体占比提高到g_c以上，那么此时传播率就可以降低到传播阈值λ_c以下。这种情况下τ为负数，病原体将自然消亡。这解释了为什么卫生官员鼓励大部分人群接种流感疫苗：疫苗不仅能保护接种的个体，还能通过降低病原体传播率来保护未接种的个体。与之类似，避孕套不仅能保护使用者，使其避免接触人类免疫缺陷病毒，也降低了艾滋病在整个性关系网络中的传播率。因此在随机网络中，足够高的免疫率可以使得病原体从人群中消亡。

- 异质网络

如果病原体在具有较高$\langle k^2 \rangle$ 的网络中传播，且随机免疫将λ变成λ(1-g)。那么根据公式**9.26** ，我们确定临界免疫率g_c 满足

$$\frac{\beta}{\mu}(1-g_c)=\frac{\langle k \rangle}{\langle k^2 \rangle} \quad (9.28)$$

边栏**9.3**

如何阻止流行病传播？

卫生安全官员依赖多种干预措施来控制或延缓流行病暴发。最常用的几种干预措施包括：

降低传播率的措施

戴口罩、手套和洗手可以降低空气传播或接触传播的病原体的传播率。与之类似，避孕套降低了性传播病原体的传播率。

减少接触的措施

对于后果严重的疾病，可以隔离病人，关闭学校，限制人们进出人流密集的公共场所（例如电影院和商场）。这减少了个体的邻居数目，导致接触网络更稀疏，从而降低了传播率。

疫苗

疫苗可以从网络中将接种节点永久移除，因为这些节点不再会被感染，也不会传播疾病。疫苗也降低了传播率，从而提高了病原体消亡的概率。

可得：

$$g_c=1-\frac{\mu}{\beta}\frac{\langle k \rangle}{\langle k^2 \rangle} \quad (9.29)$$

在随机网络中，公式**9.29** 简化成公式**9.27** 。对于γ ＜3的无标度

网络，我们有$\langle k^2\rangle \to \infty$，因此公式**9.29** 指出$g_c \to 1$。换言之，如果接触网络有很高的$\langle k^2\rangle$，我们几乎需要使所有的节点都免疫才能阻止流行病的传播。这一结果与一些实际发现是一致的：对于很多疾病，我们需要使总人口的80%~100%都免疫才能灭绝其病原体。例如，麻疹要求95%的人口接受免疫[8]，基于随机免疫的数字病毒免疫策略要求接近100%的计算机安装合适的反病毒软件[67]。

为了说明度异质性在免疫中的影响，我们假设现在有一种数字病毒正在电子邮件网络上传播。如果这一电子邮件网络是随机无向的，我们有$\langle k\rangle =3.26$。令公式**9.27** 中的$\lambda =1$可得$g_c =0.76$。换言之，要消灭这一病毒，我们需要说服76%的计算机用户更新他们的反病毒软件。但事实上，电子邮件网络是无标度的，且$\langle k^2\rangle =1\quad 271$（无向），此时公式**9.27** 不再适用。在这种情况下，公式**9.29** 指出$\lambda =1$时，$g_c =0.997$，即需要99.7%的用户安装反病毒软件才能阻止这种电子邮件病毒的传播。这几乎是一个不可能达到的比例——很多用户总是忽略所有警告。这就是为什么电子邮件病毒能存活多年，直到其生存的操作系统被淘汰[67]。

无标度网络上的接种策略

随机免疫失效的原因根植于传播阈值趋近于零。因此，为了在异质网络中成功消灭病原体，必须想办法提高传播阈值。这就要求我们尽可能地降低底层接触网络的方差$\langle k^2\rangle$。

异质网络方差大是因为存在枢纽节点。如果我们让枢纽节点（度大于预先选定的几个节点的最大值$k'_{\max}$）免疫，就可以降低网络的度方差并提高公式**9.26** 中的传播阈值[68]，[69]。事实上，如果度$k > k'_{\max}$的节点都消失，传播阈值将变为（进阶阅读**9.C**）：

$$\lambda'_c \approx \frac{\gamma-2}{3-\gamma}\frac{k_{\min}^{2-\gamma}}{(k'_{\max})^{\gamma-3}} \qquad (9.30)$$

因此，对于$\gamma <3$的情况，我们治愈/免疫的枢纽节点越多（$k'_{\max}$越小），传播阈值就越大（图**9-22**）。通过让足够多的枢纽节点

点免疫，我们得以将λ_c降到病原体的$\lambda=\beta/\mu$以下。这个过程等价于改变底层网络：通过让枢纽节点免疫，使接触网络碎片化，从而让病原体难以到达其他连通分支的节点（图**9-23**）。

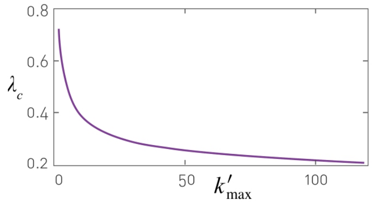


图**9-22**　免疫枢纽节点

在异质网络中，可以通过让枢纽节点免疫来提高传播阈值，以消灭病毒。图中展示了当我们让所有度大于$k'_{\max}$的节点免疫时的期望传播阈值。越多枢纽节点得到免疫（$k'_{\max}$越小），λ_c越大，越可能让这一疾病消亡。让枢纽节点免疫使得这些节点对病原体不可见，改变了疾病传播的网络（图**9-23**）。

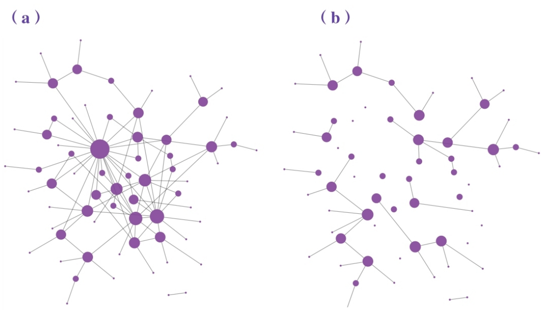


图**9-23**　鲁棒性和免疫

当节点和链接出现随机失效时，无标度网络表现出强大的韧性（第**7**章）。同时，无标度网络受到攻击时是脆弱的：如果我们移除大度节点，无标度网络就会分裂成碎片。这一现象与免疫问题有诸多相似之处：随机免疫无法消灭传染病，但有针对性地对枢纽节点免疫可以让传染病存在有限的临界阈值，从而帮助我们消灭传染病。这一相似性并不是偶然的，鲁棒性和免疫问题都与发散的$\langle k^2\rangle$有关。事实上，传播阈值归零等价于另一发现：随机移除节点后渗流阈值收敛到1（进阶阅读**9.D**）。与之类似，枢纽节点免疫后，传播阈值升高等价于无标度网络受到攻击时的低渗流阈值。因此，网络攻击和针对性免疫问题是同一个硬币的两面。为了理解网络攻击和针对性免疫的等价性，考虑图a中所示的网络。网络攻击移除了5个最大的枢纽节点，使网络分裂成许多孤立的小岛，如图b所示。针对性免疫扮演了同样的角色：通过让枢纽节点免于传染病感染，病原体传播的网络变成了图b中所示的碎片网络。随着免疫后的网络分裂成小岛，病原体只能在局部传播，无法感染其他部分的节点。

让枢纽节点免疫的策略其实也代表了关于免疫策略的观点的变化：不同于使用随机免疫来降低传播率，我们必须改变接触网络的拓扑结构，从而将λ_c提升至生物属性所决定的$\lambda = \beta / \mu$。

但基于枢纽节点的免疫策略也存在一定问题：对于大多数流行病传播过程，我们缺乏接触网络的详细数据。事实上，我们既不知道人群中每个个体的性伴侣个数，也无法在流感暴发中准确地找到超级传播者。换言之，我们很难知道哪些节点是枢纽节点，不过还是可以利用网络的拓扑性质设计一种比随机免疫更有效的免疫策略。友谊悖论告诉我们：平均而言，一个节点的邻居的度高于该节点的度（边栏**6.1**）。借助这一性质，在无法确切知道哪些节点是枢纽节点的情况下，我们可以通过随机选择一个个体并让其邻居免疫的方法尽可能地囊括它们。整个过程分成以下3个步骤[70]：

（**1**）与随机免疫类似，随机选择一部分节点（占比为p），记为第0组。

（**2**）对第0组中的每个节点，随机选择一条链接。将这些链接指向的节点记为第1组。例如，我们要求第0组中的每个人指出一个可能会相互传播病原体的朋友。在考虑HIV的情况时，则要求每个人说出一个性伴侣的名字。

（**3**）让第1组的个体免疫。

这种策略不需要知道网络的全局结构。但是根据公式**6.3**，一个有k条链接的节点属于第1组的概率正比于kp_k。因此，第1组个体的平均度将高于第0组个体。图**9-24**通过展示在无标度网络上消灭一种病原体所需的临界阈值与度指数γ的关系，说明了这一偏差导致的不同结果。该图给出了如下关键结论：

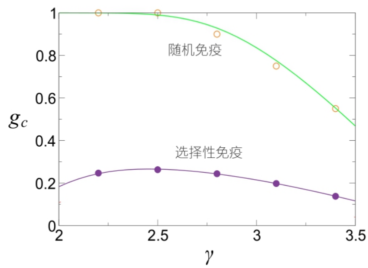


图9-24　无标度网络的选择性免疫

SIS模型中，临界免疫阈值g_c与病原体传播的接触网络的度指数γ的关系。曲线表示了两种免疫策略：随机免疫（绿色）和选择性免疫（紫色）。后者是指随机选择一个节点并让其第一个邻居免疫。连续的曲线表示解析结果，图中的点是在$N=10^6$和$m=1$的情况下模拟的结果。因为人口总数是有限的，即使在$\gamma<3$的条件下，仍有$g_c<1$。[70]

（1）随机免疫

图9-24中上方的曲线表示随机免疫的g_c。对于异质网络（γ较小），我们发现$g_c\approx 1$。这表明必须让所有节点免疫才能消灭这一传染病。当γ接近3时，网络存在有限的传播阈值，g_c降低。因此当γ较大时，让人群中足够多的个体免疫就可以消灭病原体。

（2）选择性免疫

在有偏策略中，g_c稳定地处于30%以下。因此让30%节点的一个随机选择的邻居免疫，就可以消灭这一传染病。这一策略的有效性仅在一定程度上依赖于γ。即使在γ较大时，即枢纽节点不那么重要时，选择性免疫仍比随机免疫更高效。

总而言之，如果有充足的资源，可以让每一个受到病原体威胁的人都免疫，那么我们就应该这么做——这是以前消灭传染病运动的策略（边栏9.4）。若很难做到大规模免疫，就需要采取不同的免疫策略来让我们所能使用的资源达到最大的效益。每种策略的有效性取决于病原体传播的接触网络的结构。一般而言，若病原体在异质网络中传播，则随机免疫是低效的：需要找到并免疫接近100%的易感染节点才能让随机免疫策略完全起效，而这在大多数场景下是不可能的。

与之相反，对枢纽节点进行免疫的策略有很好的效果。通过选择性免疫策略对随机节点的邻居进行免疫，我们可以在不知道准确的接触网络的情况下显著提高免疫效果。这一策略对随机网络和异质网络而言都十分有效。

边栏9.4

病原体能被消灭吗？

20世纪60年代末，天花在亚洲和非洲仍然广泛存在。1967年以前，消灭天花的策略依赖大规模接种疫苗，而这在人口密集的地区是低效的。卫生官员后来制定了基于网络的策略来阻止天花传播：他们找出并治疗那些曾与受感染个体接触过的人。这一策略使得天花成为史上第一种被正式消灭的传染病（图**9-25**）。



图**9-25**　消灭天花

拉希马·巴努（Rahima　Banu），1976年孟加拉国最后一个感染天花的病人。[71]

消灭的意思是病原体从人群中完全消失。要消灭一种传染病，卫生官员首先需要确认目标病原体没有非人类宿主，从而依靠人类疫苗可以真正消灭这种传染病。同时还需要有实际可用的有效疫苗或药物来阻断其传播。截至目前，消灭传染病运动取得了一部分成功——消灭了天花和牛瘟，但针对钩虫、疟疾和黄热病的运动都失败了。

9.7
流行病预测

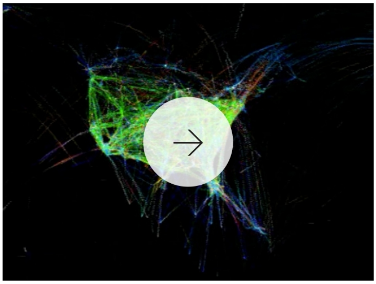
在历史的大部分时期，人类面对大规模流行病时总是束手无策。由于缺乏药物和疫苗，传染病一次次席卷全世界，夺走了大量的生命。直到1796年，才有第一支疫苗投入测试；直到20世纪90年代，针对新病原体的系统性的疫苗开发和治疗才成为可能。尽管医学发展进步巨大，我们也只是对少数病原体开发出了有效的疫苗。因此，降低传染率和基于隔离的措施仍然是卫生系统对抗新病原体的主要工具。为了让疫苗、治疗和隔离措施的组合手段生效，我们需要预测何时何地将出现下一次病原体暴发，从而让当地卫生机构优化部署可用资源。

关于流行病暴发的实时预测是近期才出现的研究领域。其发展基础来自20世纪80年代的传染病模型框架 [72] 和2003年非典型性肺炎大暴发。后者确定了全球疫情的报告准则。随后出现的与流行病相关的可用数据 [1] 为建模提供了实时输入。这些技术进展首次派上用场是在2009年H1N1甲型流感暴发期间——第一次实时预测了该流行病的传播。

新病原体的出现总是伴随着几个关键问题：

- 病原体从哪里起源？
- 新发病例会出现在哪里？
- 流行病什么时候会到达各个人口稠密地区？
- 预计有多少人会被感染？
- 该如何减缓其传播进程？
- 如何能消灭它？

如今，这些问题都可以通过强大的流行病仿真工具来回答。仿真工具使用人口学信息、移动信息（在线资源**9.4**）和流行病学数据作为输入 [73-75] 。从随机种群模型 [76-78] 到基于代理人的计算机仿真（可建模百万个体的行为和交互） [79] ，都是这些工具背后的算法。在这一节中，我们将概述这些工具的用途，指出网络科学在这些技术发展中的作用。



在线资源9.4　北美航线的模式

北美实时航线图，基于美国联邦航空管理局公布的数据。全球交通网络导致了病原体在大洲之间传播。因此，航线时刻表可以作为流行病预测的输入变量。亚伦·科布林（Aaron Koblin）制作的这个视频是纯粹的科学可视化作品，亦被艺术家们视为电子艺术。事实上，这个视频是纽约现代艺术博物馆的媒体艺术藏品。

扫描二维码下载“湛庐阅读”App，回复书名，观看北美航线图的相关视频。

实时预测

流行病预测旨在预判病原体的实时传播，即预测每个主要城市的每周受感染人数 [79] ， [80] 。基于网络科学的流行病实时预测的首个成功案例，并且得益于全球疾病传播和移动（Global Epidemic and Mobility，简称GLEAM）计算模型 [80] （图9-26，在线资源9.5），这一随机框架将全世界范围的高精度人口数据和移动数据作为输入，采用了基于网络的计算模型：

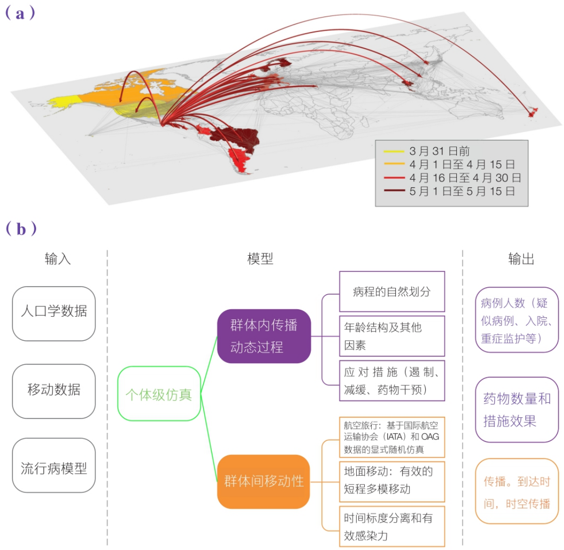


图9-26　建模2009年H1N1甲型流感传播

（a）　H1N1甲型流感病毒在2009年暴发期早期的传播。箭头表示首个感染者到达之前未被感染的国家。颜色表示病毒到达的时间。

（b）　GLEAM的计算模型流程图，用于预测H1N1甲型流感病毒和埃博拉病毒等病原体的实时传播。左栏（输入）表示输入的数据库，包括人口学信息、移动信息和流行病学信息；中栏（模型）描述了每一时间步建模的基于网络的动态过程；右栏（输出）是模型能预测的物理量的示例。[82]



在线资源9.5　GLEAM

介绍流行病预测软件包GLEAM的视频。

扫描二维码下载“湛庐阅读”App，回复书名，获取视频。

- GLEAM将每个地理位置映射为网络中的节点。
- 根据全球交通数据（例如航线时刻表），用节点之间的传输表示链接（在线资源**9.4**）。
- GLEAM使用基于网络的方法估算流行病参数，例如传播率和再生数。它基于历年全球范围流行病的传播数据，而非医学报告[81]。

GLEAM实现了**9.3**节 中描述的基于网络的流行病框架，给出未来数月内病原体在全球传播的多种可能的结果。对于H1N1甲型流感，我们将该模型的预测结果与48个国家在疫情全阶段从监测和病毒学来源与收集到的数据进行对比[80]，有如下几点关键发现：

- 高峰期

高峰期是指某个国家病原体感染人数最多的一个星期。对高峰期的预测可以帮助卫生官员决定分配疫苗或治疗手段的时间和数量。高峰期与首个病例到达的时间，以及每个国家的人口学特征和移动行为特征等有关。在87%的国家中，观测到的高峰期均在预测区间内（图**9-27**）。剩下的国家中，观测的高峰期和预测的高峰期最多相差两星期。

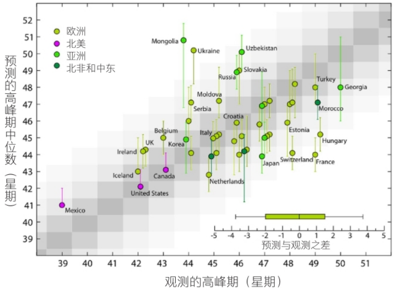


图**9-27** **H1N1**甲型流感活动高峰期

多个国家中H1N1甲型流感病毒预测和观测的活动高峰期。高峰期是指病原体感染人数最多的一个星期，以距离流行病开始的星期数标识。模型通过分析2 000次暴发的

随机实现来预测，并生成了图中的误差线。[82]

- 早到的高峰期

GLEAM预测H1N1甲型流感的高峰期将出现在11月，而不是1月或2月——流感类病毒的常规高峰期。这个意外的预测结果后来被证实是正确的，进一步证明了模型的预测能力。高峰期提前是因为H1N1甲型流感原发地在墨西哥，而不像其它很多流感那样原发于南非，因此病毒到达北半球的时间提前了。

- 接种疫苗的影响

一些国家实施了疫苗接种运动来控制这次的疫情。仿真结果表明，这些大规模疫苗接种运动对这次流行病传染过程的影响几乎可以忽略不计。原因在于，疫苗接种运动的时间是根据1月高峰期预估推算的，因此都在2009年11月之后进行[83]。而此时已经过了高峰期，很难有好的效果。

假设分析

通过考虑各种遏制、减缓措施的时间和性质，仿真可以估计各种应急方案的效果[73-75]，[77]，[84]。接下来我们将讨论两种此类干预措施的影响。

- 旅行限制

由于航空旅行在病原体传播中扮演了重要角色，当我们面临埃博拉病毒这样的危险流行病时（图**9-28**），直觉上的首选方案就是限制旅行。然而在当今世界，航空运输有关键的作用，禁止航空旅行会导致经济崩溃。因此在实施航空禁令之前，我们必须确认限制旅行确实有助于控制疫情。其中一个很重要的认识是，“意识到病毒暴发”这件事本身就会减少旅行。例如，2009年5月，H1N1甲型流感暴发期间，进出墨西哥的人流量下降了40%——人们取消了在受感染地区进行的非必要商务和休闲活动。模型指出[80]，[82]，这减少的40%使世

界上各个国家的首个病例到达时间推迟了不超过3天。我们进一步认为，即使减少90%的旅行，高峰期的推迟也不会多于20天（图**9-29**）。

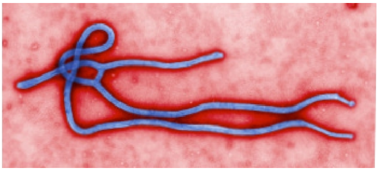


图**9-28**　致命暴发

致死率约80%的埃博拉病毒是人类已知最致命的病毒之一。它于1976年在扎伊尔首先被发现，312名感染者中有280名死于出血热（兼有高烧和出血的症状）。接触已感染个体血液或分泌物可传播这一病毒。

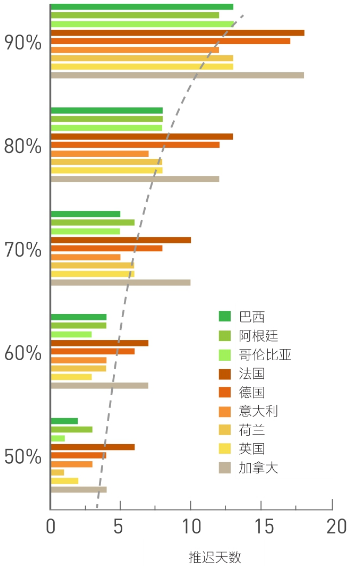


图**9-29**　减少旅行的影响

减少旅行如何影响H1N1甲型流感病毒从墨西哥到各个国家的到达时间（相对于不减少旅行的参考情况）。纵轴的百分比表示全世界范围内减少旅行的比例。在减少90%旅行的情况下，最长的延迟不超过20天。[77]

最重要的是，旅行限制并不会减少受感染人数，而只能延迟疫情暴发，让当地政府机构有更多的时间准备应对流行病。因此，只有当

旅行限制所导致的延迟能够帮助提高当地疫苗接种水平或改善治疗手段时，旅行限制才是有效的。

边栏9.5

电影之夜

电影《传染病》（*Contagion*）描述了一场大流行病，是一部看起来很真实的科幻作品。这部2011年的医学惊悚电影由史蒂文·索德伯格（Steven Soderbergh）执导，主演包括玛丽昂·歌迪亚（Marion Cotillard）、布莱恩·克兰斯顿（Bryan Cranston）、马特·达蒙（Matt Damon）、劳伦斯·菲什伯恩（Laurence Fishburne）、裘德·洛（Jude Law）、格温妮丝·帕特洛（Gwyneth Paltrow）、凯特·温斯莱特（Kate Winslet）和詹妮弗·艾莉（Jennifer Ehle）。电影描述了公共卫生官员拼命阻止病毒及其引发的恐慌席卷全球，展示了生物和社交接触的影响。1995年，医学灾难电影《极度恐慌》（*Outbreak*）关注了一种从扎伊尔的一个小村庄开始蔓延至美国的类似埃博拉的致命病毒。这部电影由沃尔夫冈·彼得森（Wolfgang Petersen）执导，主演包括达斯汀·霍夫曼（Dustin Hoffman）、蕾妮·罗素（Rene Russo）和摩根·弗里曼（Morgan Freeman）。两部电影都展示了民事和军事机构为了遏制致命病原体蔓延而必须做出的艰难抉择。



图9-30 暴发：科幻和现实

两部与流行病相关的电影的发行海报：《传染病》和《极度恐慌》。

- 抗病毒治疗

在2009年H1N1甲型流感传播时期，加拿大、德国、中国香港、日本、英国和美国都分发了抗病毒药物，来减轻这一传染病的影响[85]。这不禁使人发问：如果有药物储备的所有国家都向全部人口分发药物，会产生什么影响[86]？仿真实验表明，高峰期会推迟3至4星期，从而使人们有更多的时间，让全部人口中更大比例的人在疫情达到高峰之前获得免疫。

有效距离

在汽车和飞机发明之前，病原体以步行速度，或最多以马车的速度进行传播。因此，像欧洲的黑死病这样的大规模疫情也以缓慢的速度从一个村子传播到另一个村子（图**9-8**），其扩散过程可以用简单的反应 - 扩散模型（reaction-diffusion models）来描述[87]，[88]。由于新增感染者总是跟之前的感染者存在地理邻近性，因此暴发时间曾与暴发地之间的物理距离存在很强的相关性。

而如今，有了航空旅行之后，物理距离失去了与疫情传播的相关性。曼哈顿出现的病原体可以很容易地到达伦敦，就像到达纽约州的加里森（Garrison）一样——后者距离曼哈顿只有一小时车程。这促使我们提出了一个问题：除了物理空间之外，有没有更好的观察流行病传播的空间呢？这样的空间是有的，只要我们把传统的地理距离替换为从移动网络中推算出的有效距离即可[89]。移动网络的节点是城市，链接表示城市之间的出行量。链接是有向加权的，权值为流量比例（flux-fraction）$0 \leq p_{ij} \leq 1$，表示出行人群中离开节点i到达节点j的比例。p_{ij}的值可以从航线时刻表中抽取：仅当i到j有直接航线时，才有$p_{ij}>0$。

因为两个城市之间的路径可以有很多条，病原体也可以沿着移动网络上的多条路径传播。然而，决定实际传播的是由移动矩阵p_{ij}计算出的最大概率路径。这就可以使我们对两个相邻位置i和j之间的有效距离d_{ij}做出如下定义：

$$d_{ij} = (1 - \ln p_{ij}) \geqslant 0 \qquad (9.31)$$

如果p_{ij}较小，则只有很少比例的人离开i前往j，那么i和j之间的有效距离较大。注意$d_{ij} \neq d_{ji}$：小村庄i紧邻大城市j，我们认为d_{ij}很小，因为i中有很多人前往j。然而，d_{ji}很大，因为只有很少一部分人离开大城市前往小村庄。公式**9.31**中取对数使得有效距离可累加，对应于多步路径上的概率可连乘。

如图**9-31**所示，如果我们用公式**9.31**表示每个城市到流行病原发地的距离，则病原体以圆形波阵面（circular wave fronts）传播。相反，如果我们在物理距离空间观察流行病传播时，会看到非常复杂的传播模式。此外，H1N1甲型流感的到达时间关于物理距离的函数看起来非常随机，而其与有效距离之间却存在很强的相关性（图**9-32**）。因此，我们可以利用有效距离来确定病原体的传播速度（在线资源**9.6**）。

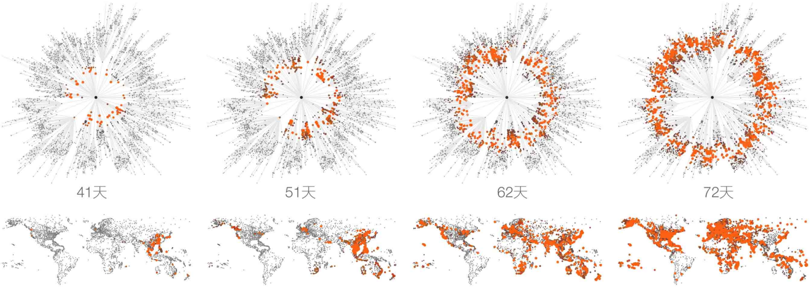


图9-31　有效距离

始发于中国香港的疫情传播。有大量感染病例的地区标记为红色，每个子图比较了以地理位置（下图）和有效距离（上图）表示的系统状态。地理位置表示出的复杂空间模式，在有效距离表示中变成了以恒定速率向外扩散的圆形波（参见在线资源**9.6**）[89]。

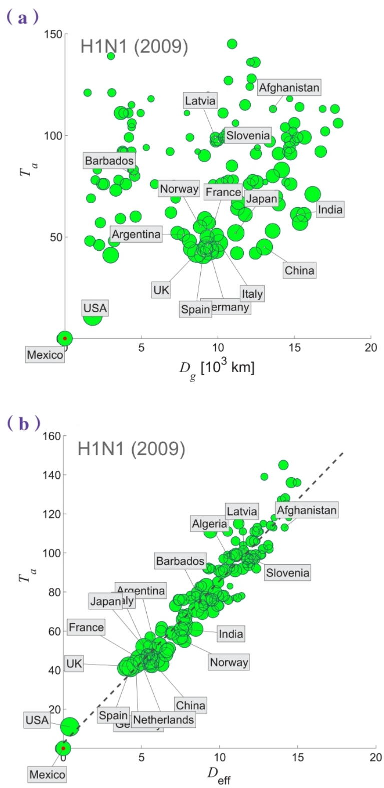


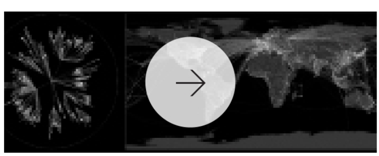
图9-32　有效距离和到达时间

（a）地理距离

2009年H1N1甲型流感期间的到达时间与原发地（也就是墨西哥）的地理距离。每个圆圈表示140个受感染国家之一，其大小代表这个国家的总交通流量。到达时间是指2009年3月17日疫情暴发后，每个国家的第一个确诊病例的日期。图中到达时间和地理距离在很大程度上是相互独立的（R_0 ＝0.039 4）。

（b）有效距离

H1N1甲型流感到达时间T_a 与有效距离D_{eff} 。图中表现出有效距离公式（公式**9.31**）到达时间的强相关关系。[89]



在线资源9.6　疫情传播速度

GLEAM预测的从3个初始暴发地区开始的疫情传播。虽然地理上的传播模式难以解释，但在用有效距离重新表示后，我们可以看到疫情传播遵循放射状模式（图**9-31**）。观察到的这一传播模式给我们提出了一个问题：一个典型的病原体在全球传播的速度是怎样的呢？速度取决于3个关键参数：

（1）基本再生数R_0。流感类病毒大约为2（表**9-2**）。

（2）康复概率。流感大约3天。

（3）移动率，即每天出行的人占总人口的比例。通常介于0.001~0.01之间。

输入这些参数运行GLEAM（图**9-26**），我们可以计算到达时间和与传染病原发地之间的地理距离的相关性，得到病毒大约每天传播250~300公里。因此，流感病毒在大洲内部的传播速度相当于跑车或者小型飞机[89]。

传染病预测问题有一个令人惊讶但喜闻乐见的方面：即使使用不同的移动数据（航线时刻表[25]，[26]或美元移动[24]）、假设不同的流行病参数（康复概率、传播率等），各种模型的估计结果也都非常接近。而有效距离可以帮助我们理解为什么不同的模型会给出相近的预测。事实上，我们可以把病原体到达位置a的时间写作[89]：

$$T_a = \frac{d_{\text{eff}}(P)}{V_{\text{eff}}(\beta, R_0, \gamma, \varepsilon)} \quad (9.32)$$

因此，到达时间等于有效距离d_{eff}和有效速度V_{eff}之比。有效速度只取决于病原体的流行病学参数，而有效距离d_{eff}只取决于移动网络的拓扑结构——链接权重p_{ij}。在出现新的流行病暴发时，一开始我们并不知道其病原体的流行病学参数。然而，公式**9.32** 说明相对到达时间是与流行病学参数无关的。例如，对于节点i 开始的流行病暴发，其到达节点j和节点l的时间之比为

$$\frac{T_a(j/i)}{T_a(l/i)} = \frac{d_{\text{eff}}(j/i)}{d_{\text{eff}}(l/i)}$$

即该比例仅取决于有效距离。因此，传染病的相对到达时间只取决于移动网络的拓扑结构。由于全世界范围内的移动模式是唯一的且与模型无关，因此不同模型的预测结果也几乎一致——与其各自选取的流行病学参数无关。

总而言之，数据的收集和网络上的传染病研究的进步共同促使实

时预测病原体传播成为可能。所开发的模型可以帮助设计应对和缓解方案，训练医疗和急救人员，还可以用于探索从隔离到旅行限制等不同干预措施的影响，优化治疗和疫苗的分配。

有趣的是，流行病预测在近期取得成功并不是因为人们对传染性病原体的生物属性有了进一步的认识。应该说这是一种很幸运的情况，当涉及病原体传播时，其流行病学参数并不是最重要的。最重要的因素是移动网络的拓扑结构。而这一结构可以从旅行日程表中准确构建出来，从而使我们可以从人类的移动模式中得到关于流行病传播的准确预测。

边栏9.6

确定流行病的源头

确定流行病的源头是流行病控制的重要部分。源头可能是接触网络的第一个个体，或者移动网络中病原体首次出现的城市。这个问题的数学表达[91]引发了大量研究[92-99]。

寻找源头的困难之处来自感染过程的随机性：不同的初始条件可能导致类似的感染模式被观察到。应对的方法取决于我们对该流行病掌握的信息：

- 最简单的情况，在给定的时刻t，我们知道已感染的节点和病原体传播的网络。这时的任务是寻找源节点i[91]（图9-33）。



图9-33　流行病的源头

寻找流行病的原发地就像寻找水波的来源。因为病原体并不在均匀介质中扩散，在移动网络中定义合适的“水波”是很困难的。

- 如果还知道每个节点的感染时间，我们可以重构该流行病传播的动态过程，从而显著提高寻找源节点的能力。
- 最佳策略是监控枢纽节点，因为它们有流行病暴发的最早信息和最准确信息。例如，若病原体在无标度网络上传播，监控18%的大度节点的状态就能以90%的正确率找到源节点。与之相反，若随机选择节点进行监控，则需要监控41%的节点才能达到同样的正确率[93]。
- 当以有效距离表达时（图**9-31**），传播过程仅在我们找到了正确的暴发位置时才遵循圆形模式。否则我们会观测到不对称的传播模式。因此，可以以这样的方式寻找：源节点应该使得传染病暴发模式呈现最大程度的径向对称[89]。

9.8 小结

大多数网络都是沿着其链接进行传播的，包括传播信任、知识、习惯或信息（社交网络）、电力（电网）、货币（金融网络）、商品（贸易网络）等。要理解这些现象，必须理解网络拓扑是如何影响其动态传播过程的。在这一章，我们关注病原体沿着网络的链接进行传播的现象。在这一领域，我们关于动态现象和网络拓扑之间相互作用的理解是最前沿的。我们发现，网络拓扑结构对动态传播的过程有极大的影响——随机网络和无标度网络上的传播会产生截然不同的结果。这一发现也为解决更广泛的一类问题奠定了基础，即系统地理解网络对不同动态过程的影响[100]。而这也成了网络科学领域越来越活跃的一个课题[101]，[102]。

建模病原体的传播也是网络科学的一个重要实践应用。这一领域所取得的进展是令人瞩目的：我们实现了对流行病的精确预测，而这在十年前还只是一个梦。这一进展得益于两方面的进步：一是出现了鲁棒的理论框架来描述基于网络的流行病；二是可以获取实时的人类

旅行和人口学数据，使我们得以重构引发病原体在全球范围内传播的移动网络。如我们在**9.7**节　所看到的，生物属性和网络结构对模型预测能力的准确性都有各自的贡献，但并非紧密耦合的。因此，如果需要一个准确的预测，首先就要对移动网络有一个准确的了解。

　　网络上的流行病分析框架给出了很多让人意想不到的结果，其中最重要的一个结果是：在异质网络上，特征传播时间和传播阈值趋近于零。由于流行病传播所涉及的接触网络大多具有较宽的度分布，这些结果同时具有短期和长期的理论与应用价值。

边栏**9.7**

快速回顾：网络上的流行病

感染率：β

康复概率：μ

传播率：$\lambda = \dfrac{\beta}{\mu}$

再生数：$R_0 = \dfrac{\beta\langle k\rangle}{\mu}$

易感染－已感染模型（**SI**模型）：

$$i(t) = \frac{i_0 e^{\beta\langle k\rangle t}}{1 - i_0 + i_0 e^{\beta\langle k\rangle t}}$$

易感染－已感染－易感染模型（**SIS**模型）：

$$i(t) = \left(1 - \frac{\mu}{\beta\langle k\rangle}\right)\frac{Ce^{(\beta\langle k\rangle-\mu)t}}{1 + Ce^{(\beta\langle k\rangle-\mu)t}}$$

特征时间：

SI模型：

$$\tau = \frac{\langle k\rangle}{\beta\left(\langle k^2\rangle - \langle k\rangle\right)}$$

SIS模型：

$$\tau = \frac{\langle k \rangle}{\beta\left(\langle k^2 \rangle - \mu\langle k \rangle\right)}$$

SIR模型：

$$\tau = \frac{\langle k \rangle}{\beta\langle k^2 \rangle - (\mu + \beta)\langle k \rangle}$$

传播阈值：

SIS模型：

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle}$$

SIR模型：

$$\lambda_c = \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$$

免疫阈值（SIS模型）：

$$g_c = 1 - \frac{\mu}{\beta}\frac{\langle k \rangle}{\langle k^2 \rangle}$$

同样重要的还有网络上的流行病学对于免疫策略所提供的支持。如我们在**9.6**节　所展示的，虽然随机免疫能够成功地消灭在随机网络上传播的病毒，但这一策略并不适用于无标度网络。而由于大多数接触网络都是异质网络，这是一个相当令人沮丧的结论。不过，我们发现选择性免疫策略可以让传播阈值回到零以上，进而阻止病原体的蔓延。这一策略的成功是因为它系统地改变了病原体传播的网络拓扑结构。

边栏9.8

史料：网络上的流行病

罗穆阿尔多·帕斯托尔 - 萨托拉斯和亚历山德罗·维斯皮那尼引入连续介质理论来解释底层接触网络性质后，流行病现象成为网络科学的核心课题。他们还发现，传播阈值和特征时间依赖于度分布的二阶矩，这是网络上的流行病研究的一个核心结论。维斯皮那尼和他的研究组开发了可用于实时预测病原体传播的计算框架GLEAM。



图**9-34**　罗穆阿尔多·帕斯托尔 - 萨托拉斯与亚历山德罗·维斯皮那尼

罗穆阿尔多·帕斯托尔 - 萨托拉斯获得物理博士学位后作为博士后与亚历山德罗·维斯皮那尼一起在的里雅斯特（Trieste，意大利北部城市）的ICTP工作。他们发现了无标度属性对于传播阈值的影响。后来，这两位研究者都对网络科学做出了重大贡献，包括发现度相关性（第**6**章 ）和理解加权网络。

9.9
课后习题

习题1　网络上的流行病模型

对于具有下列度分布的网络，分别计算SI、SIS和SIR模型的特征时间τ 和传播阈值λ_c 。

（**1**）指数分布；

（**2**）广延指数分布；

（**3**）狄拉克函数分布。

假设网络的度没有相关性且无穷大。各个分布的函数形式及对应

的一阶矩和二阶矩参见表**3-2** 。

习题2　随机分布的社交网络上的肥胖现象

考虑一个度分布为p_k 的社交网络，其中50%的节点是肥胖的。假设肥胖节点在网络中随机分布。

（**1**） 如果网络存在以联合概率$e_{kk'}$表示的度相关性，求一个非肥胖个体(ϕ)与一个肥胖个体(o)是朋友的概率P（ϕo ），以及两个肥胖个体是朋友的概率P (oo)。

（**2**） 假设网络没有度相关性。一个度为k 的节点有多少个肥胖的二阶邻居？

若肥胖比例上升到70%，（1）和（2）的答案分别是多少？

习题3　免疫

从表**3-1** 中选择4个网络（有向网络的度分布按照无向和不相关网络类似计算$p_k = p_{k_{\mathrm{in}}}$ ），考虑其上的流行病传播过程。记住：除了病原体之外，思想和观点也可以在网络上传播！如果随机免疫比例g 的节点，确定每个网络上能阻止传播所需的临界比例g_c 。如果所有度大于1 000的节点都能获得免疫，传播阈值λ_c 会如何变化？

习题4　二分网络上的流行病

考虑一个二分网络，其两类节点分别表示男性（M ）和女性（F ）。在这个网络上，病原体只能从一类节点传播到另一类节点。假设从M 节点传播到F 节点的传播率为$\beta_{M\to F}$ ，从F 节点传播到M 节点的传播率为$\beta_{F\to M}$ ，$\beta_{M\to F} \neq \beta_{F\to M}$ 。假设网络度不相关，利用度块近似理论，写出SI模型的方程。

9.10
进阶阅读9.A

流行病传播的微观模型

在9.2节 和9.3节 中，我们使用连续性方式描述了传播现象。在本节中我们将展示，其主要结果也可以通过微观模型和基于概率推理的方式来得到。这些讨论有助于理解连续性方法的来源，以及更好地理解传播现象。

流行病传播方程推导

我们首先从微观过程定义两个个体之间的交互行为，并以此推导连续性模型（公式9.3 ）[101] 。考虑一个易感染个体接触了一个已感染个体，则这名易感染个体在时间dt 内以βdt 的概率受到感染，以$(1-\beta dt)$的概率不受到感染。若这名易感染个体i 的度为k_i ，则原则上它的k_i 条链接都可能感染它。因此它不受感染的概率为$(1-\beta dt)^{k_i}$ 。最终，节点i 在dt 时间内受到感染的总概率为$1-(1-\beta dt)^{k_i}$ ，即1减去其不受感染的总概率。假设$\beta dt \ll 1$，在舍去高阶项后，一名易感染个体被感染的概率约为：

$$1-(1-\beta dt) \approx \beta k_i dt \qquad (9.33)$$

在随机网络上，所有节点都有大约$\langle k\rangle$ 个邻居。将公式9.33 中的k_i 替换成$\langle k\rangle$ ，可以得到连续性方程9.3 中的第一项。如果我们不将k_i 替换成$\langle k\rangle$ ，会得到刻画病原体在异质网络上传播的公式9.13 中的第一项。

传播阈值和网络拓扑结构

9.3节 的一个关键结论是，根据连续性理论可以推导出网络拓扑结构和传播阈值λ_c 的联系。我们可以用基于微观个体的机制得到同样的结论。

考虑一个病原体在单位时间内以概率β 进行传播。那么一个度为k 的已感染节点，在单位时间内可以感染βk 个邻居。若已感染节点以概

率μ 康复，则一个节点保持已感染状态的特征时间为$1/\mu$ 。只有当已感染节点能在$1/\mu$ 的时间内感染至少一个节点时，病原体才可以长期存在，否则该病原体将会消亡。

换言之，若$\beta k/\mu<1$，则这个度为k 的节点会在成功感染其他节点之前就康复。如果我们考虑一个随机网络，其中大部分节点的度都相差不大，即$k\sim\langle k\rangle$ ，则可以根据$\beta k/\mu=1$计算传播阈值。根据$\lambda=\beta/\mu$ 可得$\lambda_c=1/\langle k\rangle$ ，这是随机网络中的公式**9.25** 在k 非常大时的极限情况。这说明病原体的传播能力由病原体的传播特性（β 和μ ）与网络拓扑（$\langle k\rangle$ ）的相互作用决定。

在无标度网络中，节点的度可以存在很大差异。此时，虽然网络的平均度可能满足让一种病毒灭绝的条件：$\beta\langle k\rangle/\mu<1$，但对于$k>\langle k\rangle$ 的所有节点，我们有$\beta k/\mu>1$。若一个满足此条件的大度节点被感染，即使传播率λ 低于阈值$1/\langle k\rangle$ ，该传染病依然可以进行传播并在枢纽节点中长期存在。这就是为什么当网络的$\langle k^2\rangle$ 较大时，传播阈值会趋近于零。

9.11
进阶阅读9.B
SI、SIS和SIR模型的解析解

在本节我们将对网络中的SI、SIS、SIR模型求解，推导出表**9-3** 列出的结果，即每种模型的特征传播时间τ 和传播阈值λ_c 。

密度函数

密度函数Θ_k 描述了度为k 的易感染节点，其邻居节点中已感染节点所占的比例。如**9.3**节 所述，要计算i_k 首先需要计算Θ_k 。若网络中不存在度相关性，则存在一条链接从一个度为k 的节点指向一个度为k'的节点的概率与k 无关。因此一条随机选择的链接，其指向一个度为k'的节点的概率为公式**6.3** ：

$$\frac{k'p_{k'}}{\sum_k kp_k} = \frac{k'p_{k'}}{\langle k \rangle}$$

每个已感染节点至少有一条链接（传递感染的链接）指向另一个已感染节点。因此可用于继续传播病原体的链接数为(k ′-1)，于是可得

$$\Theta_k = \frac{\sum_{k'}(k'-1)p_{k'}i_{k'}}{\langle k \rangle} = \Theta \qquad (9.34)$$

换言之，在没有度相关性的情况下，Θ_k 与k 独立。对公式**9.34**求微分可得

$$\frac{d\Theta}{dt} = \sum_k \frac{(k-1)p_k}{\langle k \rangle}\frac{di_k}{d} \qquad (9.35)$$

为了取得更大的进展，接下来我们需要考虑病原体传播的具体模型。

SI模型

根据公式**9.13** 和公式**9.35** 可得

$$\frac{d\Theta}{dt} = \beta\sum_k \frac{(k^2-k)p_k}{\langle k \rangle}[1-i_k]\Theta \qquad (9.36)$$

为了预测传染病传播的早期行为，我们考虑到这样一个事实，即t较小时已感染占比远远小于1。因此可以忽略公式**9.36** 的二阶项，得到

$$\frac{d\Theta}{dt} = \beta\left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)\Theta \qquad (9.37)$$

该方程的解为

$$\Theta(t) = Ce^{t/\tau} \qquad (9.38)$$

其中

$$\tau = \frac{\langle k \rangle}{\beta\left(\langle k^2 \rangle - \langle k \rangle\right)} \quad (9.39)$$

根据初始条件，即最初i_0比例的节点被均匀感染（因此对于所有的k存在$i_k(t=0)=i_0$）：

$$\Theta(t=0) = C = i_0 \frac{\langle k \rangle - 1}{\langle k \rangle}$$

可得随时间变化的密度函数Θ如下：

$$\Theta(t) = i_0 \frac{\langle k \rangle - 1}{\langle k \rangle} e^{t/\tau} \quad (9.40)$$

将其代入公式**9.13**，可得公式**9.15**。

SIR模型

在SIR模型中，已感染节点的密度为

$$\frac{di_k}{dt} = \beta(1 - i_k - r_k)k\Theta - \mu i_k \quad (9.41)$$

其中r_k是度为k的节点中已康复节点的占比。由于上式括号中的i_k和r_k在t较小时远远小于1，因此可以忽略。此时可得

$$\frac{di_k}{dt} = \beta k\Theta - \mu i_k \quad (9.42)$$

两边同乘$(k-1)p_k/\langle k \rangle$并对$k$求和可得

$$\frac{d\Theta}{dt} = \left(\beta\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} - \mu\right)\Theta \quad (9.43)$$

解公式**9.43**得

$$\Theta(t) = Ce^{t/\tau} \quad (9.44)$$

其中SIR模型的特征时间为

$$\tau = \frac{\langle k\rangle}{\beta\langle k^2\rangle - \langle k\rangle(\beta + \mu)} \quad (9.45)$$

仅当$\tau > 0$时有可能出现全球性疫情暴发，即已感染节点个数随时间呈指数增长。据此我们可以推导出全球性疫情暴发的条件为

$$\lambda_c = \frac{\beta}{\mu} > \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle} \quad (9.46)$$

从而可以得到表**9-3** 中SIR模型的传播阈值

$$\lambda_c = \frac{1}{\frac{\langle k^2\rangle}{\langle k\rangle} - 1} \quad (9.47)$$

SIS模型

在SIS模型中，已感染节点的密度由公式**9.18** 给出：

$$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta - \mu i_k \quad (9.48)$$

SIS模型的密度函数存在一个细微但很重要的差别。对于SI和SIR模型，若一个节点被感染，则它至少有一个邻居处于已感染或已康复状态。那么剩下最多有(k -1)个邻居是易受感染的，这就是公式**9.34** 括号中的（-1）项来源。然而，在SIS模型中，曾经被感染的节点可以再次变成易感染状态。因此，一个节点的k 条链接都有可能传播病原体。从而我们修改公式**9.34** 的定义，可得

$$\Theta_k = \frac{\sum_{k'} k' p_{k'} i_{k'}}{\langle k\rangle} = \Theta \quad (9.49)$$

同样，只保留一阶项可得

$$\frac{di_k}{dt} = \beta k \Theta - \mu i_k \quad (9.50)$$

两边同乘$(k-1)p_k/\langle k\rangle$并对$k$求和可得

$$\frac{d\Theta}{dt} = \left(\beta\frac{\langle k^2\rangle}{\langle k\rangle} - \mu\right)\Theta \quad (9.51)$$

同样，可以解得

$$\Theta(t) = Ce^{t/\tau} \quad (9.52)$$

其中，SIS模型的特征时间为

$$\tau = \frac{\langle k\rangle}{\beta\langle k^2\rangle - \langle k\rangle\mu} \quad (9.53)$$

当$\tau > 0$时，有可能出现全球性疫情暴发，据此可推导出全球性疫情暴发的条件为

$$\lambda = \frac{\beta}{\mu} > \frac{\langle k\rangle}{\langle k^2\rangle} \quad (9.54)$$

因此，SIS模型的传播阈值为（表**9-3**）：

$$\lambda_c = \frac{\langle k\rangle}{\langle k^2\rangle} \quad (9.55)$$

9.12
进阶阅读9.C
针对性免疫

本节中，我们将在对枢纽节点进行免疫后，分别推导SIS和SIR模型在无标度网络上的传播阈值。想象一个不相关的网络，假设其度分布服从幂律分布$p_k = c \cdot k^{-\gamma}$，其中$c \approx (\gamma - 1)/k_{\min}^{-\gamma+1}$且$k \geq k_{\min}$。在**9.3**节中，我们得到了临界传播率

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} = \frac{1}{\kappa} \text{（SIS 模型）}$$

和

$$\lambda_c = \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1} = \frac{1}{\kappa - 1} \text{（SIR 模型）}$$

免疫枢纽节点即为免疫所有度大于k_0的节点。从流行病的角度看，这等价于从网络中移除了这些大度节点。此时，为了计算临界传播率，我们需要确定移除枢纽节点后的网络平均度$\langle k' \rangle$和二阶矩$\langle k'^2 \rangle$。我们已经在进阶阅读**7.F**中讨论网络受到攻击的鲁棒性时解决了这个问题。可以看到，移除枢纽节点有两个效果：

（**1**）网络中的最大度变成了k_0。

（**2**） 与枢纽节点相关的链接也被移除了。这等价于我们随机删除如下比例的链接：

$$\tilde{f} = \left(\frac{k_0}{k_{\min}} \right)^{-\gamma+2} \quad \textbf{(9.56)}$$

此时，网络的度分布变为

$$p'_{k'} = \sum_{k=k_{\min}}^{k_0} \binom{k}{k'} \tilde{f}^{k-k'} (1 - \tilde{f})^{k'} p_k$$

根据公式**7.39**和**7.40**可以得到

$$\langle k' \rangle = (1 - \tilde{f}) \langle k \rangle$$
$$\langle k'^2 \rangle = (1 - \tilde{f})^2 \langle k^2 \rangle + \tilde{f}(1 - \tilde{f}) \langle k \rangle$$

其中，$\langle k \rangle$和$\langle k^2 \rangle$分别是移除链接前的度分布的均值和二阶矩。对于SIS模型，这意味着

$$\lambda'_c = \frac{(1 - \tilde{f}) \langle k \rangle}{(1 - \tilde{f})^2 \langle k^2 \rangle + \tilde{f}(1 - \tilde{f}) \langle k \rangle} = \frac{1}{(1 - \tilde{f})k + \tilde{f}} \quad \textbf{(9.57)}$$

其中，根据公式7.47 ，对于2>γ >3的情况，有

$$k = \frac{\gamma - 2}{3 - \gamma} k_0^{3-\gamma} k_{\min}^{\gamma-2} \qquad (9.58)$$

结合公式9.56 、9.57 和9.58 可得

$$\lambda_c' = \left[\frac{\gamma - 2}{3 - \gamma} k_0^{3-\gamma} k_{\min}^{\gamma-2} - \frac{\gamma - 2}{3 - \gamma} k_0^{5-2\gamma} k_{\min}^{2\gamma-4} + k_0^{2-\gamma} k_{\min}^{\gamma-2} \right]^{-1} \qquad (9.59)$$

对于SIR模型，类似的计算可得

$$\lambda_c' = \left[\frac{\gamma - 2}{3 - \gamma} k_0^{3-\gamma} k_{\min}^{\gamma-2} - \frac{\gamma - 2}{3 - \gamma} k_0^{5-2\gamma} k_{\min}^{2\gamma-4} + k_0^{2-\gamma} k_{\min}^{\gamma-2} - 1 \right]^{-1} \qquad (9.60)$$

对于SIR和SIS模型，若$k_0 \gg k_{\min}$ ，都有

$$\lambda_c' = \frac{3 - \gamma}{\gamma - 2} k_0^{\gamma-3} k_{\min}^{2-\gamma} \qquad (9.61)$$

9.13
进阶阅读9.D
SIR模型和键渗流

SIR模型刻画了传染病在网络中随时间动态变化的传播过程。其实它也可以映射为一个静态的键渗流问题 [103-106] 。这种映射为我们预测模型的行为提供了一套分析工具。

考虑网络中的一次流行病传播过程，每个已感染节点将病原体以概率β 传播给它的邻居，并经过康复时间$\tau = 1/\mu$ 后康复。我们将这一传染过程视为一系列随机接触行为构成的泊松过程，其中随机接触的平均时间间隔为$\beta\tau$ 。因此，一个已感染节点未传播病原体给易感染邻居的概率随时间呈指数衰减，即$e^{-\beta\tau}$ 。由于已感染节点将在康复时间$\tau = 1/\mu$ 之前一直处于已感染状态，因此病原体传播的总概率为1-e-

$\beta\tau$ 。

这个过程等价于同一网络中的键渗流过程，其中每条有向链接存在的概率为p_b ＝1-e$^{-\beta\tau}$ （图9-35 ）。若每个节点有相同的β 和τ ，则可以视网络为无向图。这样的映射虽然丢失了流行病传播过程的时序性，但仍有若干优点：

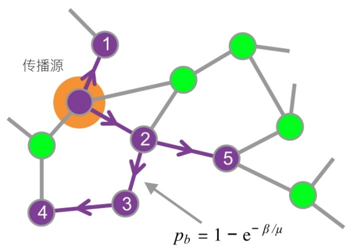


图9-35 传染病传播到渗流的映射

考虑传染病传播的接触网络。为了将传播过程映射到渗流，我们给每条链接赋予一个由病原体的生物特性决定的概率p_b ＝1-e$^{-\beta/\mu}$ 。相应地，该链接被删除的概率为e$^{-\beta/\mu}$ 。由此产生的网络的簇大小的分布可以精确对应暴发的大小。若β /μ 较大，则可能有一个巨连通分支，表明我们将面临全球性暴发。而较小的β /μ 对应难以传播的病毒，我们会观察到大量小的簇，表明病原体将消亡。

- 当流行病到达“地方病”平衡点时，已感染节点的总占比等价于渗流问题中巨连通分支的大小。
- 病原体在到达“地方病”平衡点前就消亡的概率等价于渗流问题中随机选择一个有限大小的连通分支的节点占比。
- 我们可以利用键渗流的已知特性来确定流行病的传播阈值。当选择一条链接时，考虑这条链接可到达节点的平均出度。这能够让我们追溯流行病的传播过程：如果一个已感染个体平均能感染至少一个个体，那么流行病的传播将到达“地方病”平衡点。由于一个节点可以通过其k 条链接中的任意一条被访问到，因此访问该节点的概率是kp_k /$N\langle k\rangle$ 。而它通过剩余的k -1条指出链接影响其邻居的概率为p_b 。

由于网络是随机连通的，当传染病尚未大面积扩散时，一个给定节点所感染的平均邻居数为

$$\langle R_i \rangle = p_b \sum \frac{p_k k(k-1)}{\langle k \rangle}$$

仅当$\langle R_i \rangle$ >1的时候会到达“地方病”平衡点，因此可得达到这一状态的条件为 [107] ，[108]

$$\left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right) > \frac{1}{p_b} \qquad (9.62)$$

公式**9.62** 与此前根据动态方程推导出的公式**9.46** 结果一致：对于γ ≤3的无标度网络，其二阶矩发散，因此即使p_b →0，仍存在渗流相变。也就是说，无论传染概率β 有多低、康复时间τ 有多短，病毒都能在这个网络上传播。

9.14 参考文献

[1] D. Normile. The Metropole, Superspreaders and Other Mysteries. *Science* , 339:1272-1273, 2013.

[2] J.O. Lloyd-Smith, S.J. Schreiber, P.E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature* , 438:355-359, 2005.

[3] M. Hypponen. Malware Goes Mobile. *Scientific American* , 295:70, 2006.

[4] P. Wang, M. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science* , 324:1071-1076, 2009.

[5] E. M. Rogers. *Diffusion of Innovations* . Free Press, 2003.

[6] T.W. Valente. *Network models of the diffusion of innovations* . Hampton Press, Cresskill, NJ, 1995.

[7] History and Epidemiology of Global Smallpox Eradication From the training course titled "Smallpox: Disease, Prevention,and Intervention". The CDC and the World Health Organization. Slides 16-17.

[8] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control* . Oxford University Press,Oxford, 1992.

[9] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters* , 86:3200-3203, 2001.

[10] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E* , 63:066117, 2001.

[11] Y. Wang, D. Chakrabarti, C, Wang, and C. Faloutsos. Epidemic spreading in real networks: an eigenvalue viewpoint. Proceedings of 22nd International Symposium on Reliable Distributed Systems, pg. 25-34, 2003.

[12] R. Durrett. Some features of the spread of epidemics and information on a random graph. *PNAS* , 107:4491-4498, 2010.

[13] S. Chatterjee and R. Durrett. Contact processes on random graphs with power law degree distributions have critical value 0. *Ann. Probab* ., 37: 2332-2356, 2009.

[14] C Castellano, and R Pastor-Satorras. Thresholds for epidemic spreading in networks. *Physical Review Letters* , 105:218701, 2010.

[15] B. Lewin. (ed.), Sex i Sverige. *Om sexuallivet i Sverige 1996 [Sex in Sweden. On the Sexual Life in Sweden 1996]* .National Institute of Public Health, Stockholm, 1998.

[16] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley,and

Y. Åberg. The web of human sexual contacts. *Nature* , 411:907-8, 2001.

[17] A. Schneeberger, C. H. Mercer, S. A. Gregson, N. M.Ferguson, C. A. Nyamukapa, R. M. Anderson, A. M. Johnson, and G. P. Garnett. Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sexually Transmitted Diseases* , 31: 380-387, 2004.

[18] W. Chamberlain. *A View from Above* . Villard Books,New York, 1991.

[19] R. Shilts. *And the Band Played On* . St. Martin's Press,New York, 2000.

[20] P. S. Bearman, J. Moody, and K. Stovel. Chains of affection:the structure of adolescent romantic and sexual networks. *Am J Sociol* ., 110:44-91, 2004.

[21] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature* , 453:779-782, 2008.

[22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science* , 327:1018-1021, 2010.

[23] F. Simini, M. González, A. Maritan, and A.-L. Barabási.A universal model for mobility and migration patterns. *Nature* , 484:96-100, 2012.

[24] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature* , 439:462-465, 2006.

[25] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS* , 103:2015, 2006.

[26] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *PNAS* , 101:15124,

2004.

[27] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS* , 102:7794, 2005.

[28] C. Cattuto, et al. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLOS ONE* , 5:e11596, 2010.

[29] L. Isella, C. Cattuto, W. Van den Broeck, J. Stehle, A.Barrat, and J.-F. Pinton. What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* , 271:166-180, 2011.

[30] K. Zhao, J. Stehle, G. Bianconi, and A. Barrat. Social network dynamics of face-to-face interactions. *Physical Review E* , 83:056109, 2011.

[31] J. Stehlé, N. Voirin, A. Barrat, C Cattuto, L. Isella, J-F.Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE* , 6:e23176, 2011.

[32] B.N. Waber, D. Olguin, T. Kim, and A. Pentland. Understanding Organizational Behavior with Wearable Sensing Technology. Academy of Management Annual Meeting. Anaheim, CA. August, 2008.

[33] L. Wu, B.N. Waber, S. Aral, E. Brynjolfsson, and A. Pentland. Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task. In Proceedings of the International Conference on Information Systems. Paris, France. December 14-17 2008.

[34] M. Salathé, M. Kazandjievab, J.W. Leeb, P. Levisb,

M.W.Feldmana, and J.H. Jones. A high-resolution human contact network for infectious disease transmission. *PNAS* , 107:22020-22025, 2010.

[35] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature* , 435:207-11, 2005.

[36] V. Sekara, and S. Lehmann. Application of network properties and signal strength to identify face-to-face links in an electronic dataset. *Proceedings of CoRR* , 2014.

[37] S. Eubank, H. Guclu, V.S.A. Kumar, M.V. Marathe,A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature* , 429:180-184, 2004.

[38] H. Ebel, L-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E* , 66:035103, 2002.

[39] M. Morris, and M. Kretzschmar. Concurrent partnerships and transmission dynamics in networks. *Social Networks* , 17:299-318, 1995.

[40] N. Masuda and P. Holme. Predicting and controlling infectious diseases epidemics using temporal networks. *F1000 Prime Rep* ., 5:6, 2013.

[41] P. Holme, and J. Saramäki. Temporal networks. *Physics Reports* , 519:97-125, 2012.

[42] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: how network topology and burstiness slow down spreading. *Physical Review E* , 83:025102(R), 2011.

[43] A. Vazquez, B. Rácz, A. Lukács, and A.-L. Barabási. Impact of non-Poissonian activity patterns on spreading processes. *Physical Review Letters* , 98:158702, 2007.

[44] A. Vázquez, J.G. Oliveira, Z. Dezsö, K.-I. Goh, I.

Kondor,and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E* , 73:036127, 2006.

[45] A.V. Goltsev, S.N. Dorogovtsev, and J.F.F. Mendes. Percolation on correlated networks. *Physical Review E* ., 78:051105, 2008.

[46] P. Van Mieghem, H. Wang, X. Ge, S. Tang and F. A.Kuipers. Influence of assortativity and degree-preserving rewiring on the spectra of networks. *The European Physical Journal B* , 76:643, 2010.

[47] M. Boguná, R. Pastor-Satorras, and A. Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical Review Letters* , 90:028701, 2003.

[48] Y. Moreno, J. B. Gómez, and A.F. Pacheco. Epidemic incidence in correlated complex networks. *Physical Review E* , 68:035103, 2003.

[49] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabó, D. Lazer,K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS* , 104:7332, 2007.

[50] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology* , 78:1360-1379, 1973.

[51] A. Galstyan, and P. Cohen. Cascading dynamics in modular networks. *Physical Review E* , 75:036109, 2007.

[52] J. P. Gleeson. Cascades on correlated and modular random networks. *Physical Review E* , 77:046117, 2008.

[53] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PLOS ONE* , 7:e29358, 2012.

[54] L. Weng, F. Menczer and Y.-Y. Ahn. Virality Prediction and Community Structure in Social Networks. *Scientific Reports* , 3:2522, 2013.

[55] S. Aral, and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* , 57:1623-1639, 2011.

[56] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Trans* . Web, 1, 2007.

[57] L. Weng, A Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports* , 2:335, 2012.

[58] J. Berger, and K. L. Milkman. What makes online content viral? *Journal of Marketing Research* , 49:192-205, 2009.

[59] S. Jamali, and H. Rangwala. Digging digg: Comment mining, popularity prediction and social network analysis. Proc. Intl. Conf. on Web Information Systems and Mining (WISM), 32-38, 2009.

[60] G. Szabó and, B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM* , 53:80-88, 2010.

[61] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. Proc. IEEE Intl. Conf. on Social Computing, 177-184, 2010.

[62] D. Centola. The spread of behavior in an online social network experiment. *Science* , 329:1194-1197, 2010.

[63] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth,and evolution. Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining, 44-54, 2006.

[64] M. Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology* , 83:1420-1443, 1978.

[65] N.A. Christakis, and J.H. Fowler. The Spread of Obesity in a Large Social Network Over 32 Years. *New England Journal of Medicine* , 35:370- 379, 2007.

[66] N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine* , 358:2249-2258, 2008.

[67] R. Pastor-Satorras, and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach* .Cambridge University Press, Cambridge, 2007.

[68] Z. Dezs and A-L. Barabási. Halting viruses in scale-free networks. *Physical Review E* , 65:055103, 2002.

[69] R. Pastor-Satorras and A. Vespignani. Immunization of complex networks. *Physical Review E* , 65:036104, 2002.

[70] R. Cohen, S. Havlin, and D. ben-Avraham. Efficient Immunization Strategies for Computer Networks and Populations. *Physical Review Letters* , 91:247901, 2003.

[71] F. Fenner et al. *Smallpox and its Eradication* . WHO,Geneva, 1988. http://www.who.int/features/2010/smallpox/en/

[72] L. A. Rvachev, and I. M. Longini Jr. A mathematical model for the global spread of influenza. *Mathematical Biosciences* , 75:3-22, 1985.

[73] A. Flahault, E. Vergu, L. Coudeville, and R. Grais. Strategies for containing a global influenza pandemic. *Vaccine* , 24:6751-6755, 2006.

[74] I. M. Longini Jr, M. E. Halloran, A. Nizam, and Y. Yang. Containing pandemic influenza with antiviral agents. *Am.*

J.Epidemiol ., 159:623-633, 2004.

[75] I.M. Longini Jr, A. Nizam, S. Xu, K. Ungchusak, W.Hanshaoworakul, D. Cummings, and M. Halloran. Containing pandemic influenza at the source. *Science* , 309:1083-1087, 2005.

[76] V. Colizza, A. Barrat, M. Barthélemy, A.-J. Valleron, and A. Vespignani. Modeling the world-wide spread of pandemic influenza: baseline case and containment interventions. *PLOS Med* , 4:e13, 2007.

[77] T. D. Hollingsworth, N.M. Ferguson, and R.M. Anderson. Will travel restrictions control the International spread of pandemic influenza? *Nature Med* ., 12:497-499, 2006.

[78] C.T. Bauch, J.O. Lloyd-Smith, M.P. Coffee, and A.P.Galvani. Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future. *Epidemiology* , 16:791-801, 2005.

[79] I. M. Hall, R. Gani, H.E. Hughes, and S. Leach. Real-time epidemic forecasting for pandemic influenza. *Epidemiol Infect* ., 135:372-385, 2007.

[80] M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan,B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine,* 10:165, 2012.

[81] D. Balcan, H. Hu, B. Gonçalves, P. Bajardi, C. Poletto, J.J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck,V. Colizza, and A. Vespignani. Seasonal transmission potential and activity peaks of the new influenza A/H1N1: a Monte Carlo likelihood analysis based on human mobility. *BMC Med* ., 7:45, 2009.

[82] P. Bajardi, et al. Human Mobility Networks, Travel

Restrictions, and the Global Spread of 2009 H1N1 Pandemic. *PLOS ONE* , 6:e16591, 2011.

[83] P.Bajardi, C. Poletto, D. Balcan, H. Hu, B. Gonçalves, J. J.Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V.Colizza, and A. Vespignani. Modeling vaccination campaigns and the Fall/Winter 2009 activity of the new A/H1N1 influenza in the Northern Hemisphere. *EHT Journal* , 2:e11, 2009.

[84] M.E. Halloran, N. M. Ferguson, S. Eubank, I.M. Longini,D. A. T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T.C.Germann, D. Wagener, R. Beckman, K. Kadau, C. Macken, D.S.Burke, and P. Cooley. Modeling targeted layered containment of an influenza pandemic in the United States. *PNAS* , 105:4639-44, 2008.

[85] G. M. Leung, A. Nicoll. Reflections on Pandemic (H1N1)2009 and the international response. *PLOS Med* , 7:e1000346, 2010.

[86] A.C. Singer, et al. Meeting report: risk assessment of Tamiflu use under pandemic conditions. *Environ Health Perspect* ., 116:1563-1567, 2008.

[87] R. Fisher. The wave of advance of advantageous genes. *Ann. Eugen* ., 7:355-369, 1937.

[88] J. V. Noble. Geographic and temporal development of plagues. *Nature* , 250:726-729, 1974.

[89] D. Brockmann and D. Helbing. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science* , 342:1337-1342, 2014.

[90] J. S. Brownstein, C. J. Wolfe, and K. D. Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLOS Med* , 3:e40, 2006.

[91] D. Shah and T. Zaman, in SIGMETRICS'10, Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems, pp. 203-214, 2010.

[92] A. Y. Lokhov, M. Mezard, H. Ohta, L. Zdeborová. Inferring the origin of an epidemy with dynamic message-passing algorithm. *Phys. Rev E* , 90:012801, 2014.

[93] P. C. Pinto, P. Thiran, M. Vetterli. Locating the Source of Diffusion in Large-Scale Networks. *Physical Review Letters* , 109:068702, 2012.

[94] C. H. Comin and L. da Fontoura Costa. Identiying the starting point of a spreading process in complex networks. *Phys. Rev. E* , 84:056105, 2011.

[95] D. Shah and T. Zaman. Rumors in a Network: Who's the Culprit? *IEEE Trans. Inform. Theory* , 57:5163, 2011.

[96] K. Zhu and L. Ying. Information source detection in the SIR model: A sample path based approach. *Information Theory and Applications Workshop* (*ITA*); 1-9, 2013.

[97] B. A. Prakash, J. Vreeken, and C. Faloutsos. Spotting culprits in epidemics: How many and which ones? ICDM'12;Proceedings of the IEEE International Conference on Data Mining, 11:20, 2012.

[98] V. Fioriti and M. Chinnici. Predicting the sources of an outbreak with a spectral technique. *Applied Mathematical Sciences* , 8:6775-6782, 2012.

[99] W. Dong, W. Zhang and C.W. Tan. Rooting out the rumor culprit from suspects. *Proceedings of CoRR* , 2013.

[100] B. Barzel, and A.-L. Barabási. Universality in network dynamics. *Nature Physics* , 9:673, 2013.

[101] A. Barrat, M. Barthélemy and A. Vespignani. *Dynamical*

Processes on Complex Networks . Cambridge University Press, 2012.

[102] S. N. Dorogovtsev, A.V. Goltsev, and J. F. F. Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics* , 80, 1275, 2008.

[103] R. Cohen and S. Havlin. *Complex Networks - Structure,Robustness and Function* . Cambridge University Press, 2010.

[104] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences* , 63:157, 1983.

[105] M. E. J.Newman. The spread of epidemic disease on networks. *Physical Review E* , 66:016128, 2002.

[106] C. P. Warren, L. M. Sander, and I. M. Sokolov. Firewalls,disorder, and percolation in networks. *Mathematical Biosciences* , 180:293, 2002.

[107] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin. Resilience of the Internet to random breakdown. *Physical Review Letters* , 85:4626-4628, 2000.

[108] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J.Watts. Network robustness and fragility: percolation on random graphs. *Physical Review Letters* , 85:5468-5471, 2000.

结语
运用网络科学理解当今的互联世界

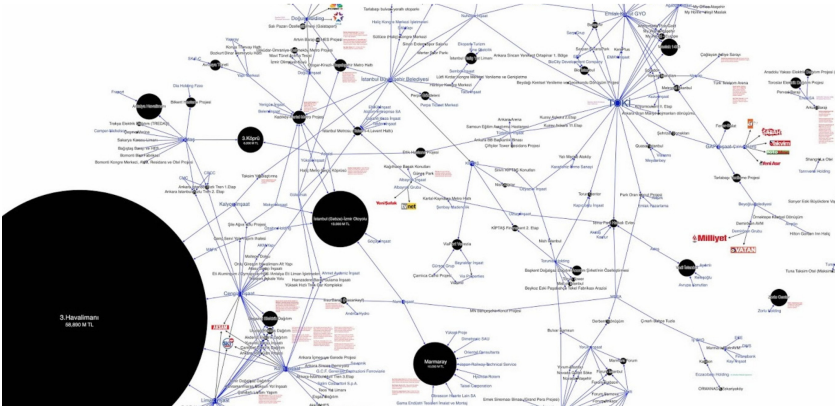


图10-0　所有权网络

2014年，一场大规模的抗议运动在土耳其轰动一时，数千名激进分子和抗议者在盖齐公园（Gezi　Park）集会。这场抗议运动还伴随着网络空间的抗议活动——抗议者通过Twitter和万维网动员支持者。这场运动的核心可以通过所有权网络来刻画，该网络由艺术家、律师、激进分子和记者联盟提供，描绘了伊斯坦布尔政商两界精英背后复杂的财务关系。这幅图于2013年在伊斯坦布尔双年展上首次展出，我们在图中使用黑色圆圈标记项目的所有权，圆圈的大小表示该项目的货币价值；用蓝色来标识企业和新闻媒体，它们和项目之间直接相连。职务犯罪标记为红色，土耳其申办奥运会的支持者标记为紫色，伊斯坦布尔双年展的赞助商标记为绿色。该图由亚萨尔·阿达纳里（Yaşar Adanalı）、布拉克·阿勒坎（Burak Arıkan）、奥兹古尔·森亚（Özgül Şen）、赛诺·乌斯塔曼（Zeyno　Üstün）、奥兹莱姆·兹戈兰德（Özlem　Zıngıland）和一些匿名参与者使用绘图软件Graph Commons绘制。

本章索引

10.1　如何使用本书 / **543**

10.2　致谢 / **549**

本章致谢

妮科尔·萨迈　**NICOLE SAMAY**

萨拉·莫里森　**SARAH MORRISON**

10.1 如何使用本书

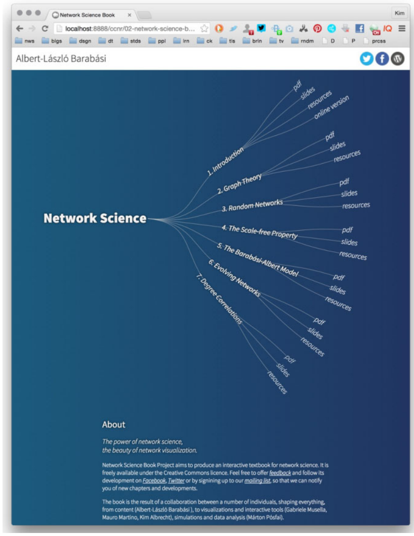
网络视角对于理解当今这个互联的世界是不可或缺的。这本书为了解网络视角提供了一个通道，为每个人提供了一个成为网络科学家的机会。在选择本书的主题和表述方式时，我力求提供一种定量的、容易理解的方式，使读者能够在愉悦的阅读体验中了解网络科学。同时，我试图向读者传递网络为我们理解身边众多复杂系统所提供的洞察力。为了解决上述两个诉求之间的矛盾，我将技术进步和一些历史注释、专题结合起来，讲述了关键科学发现的起源和应用。

写这篇结语有两个目的。一方面，我将介绍如何使用这本书；另一方面，我想通过这篇结语向为本书提供帮助的人和单位表示感谢。

线上课堂

网络科学的很多内容和知识，最适合在线展示。因此，在本书的章节中，我们会碰到很多指向在线资料的链接——“在线资源”，包括视频、软件、交互工具、数据集和数据源。要使用这些资源，请访问网站：http://barabasi.com/NetworkScienceBook。

网站上还提供了我在讲授网络科学课程时使用过的一些课件，这些课件和本书的内容相呼应。任何讲授网络科学的人都可以自由使用这些课件，并可以根据各自的授课经验修改它们。出于教学目的使用这些课件，则不需要向作者申请许可。



在线资源10.1　**barabasi.com/NetworkScienceBook**

该网站中有本书的电子版、相关视频、在前述章节中的在线资源部分提到的软件和交互工具、我讲述网络科学课程时的课件和本书中分析过的数据集。

网络科学源于实证，因此本书特别强调对真实网络的分析。为此，我们收集了10个经常在网络科学的文献中被使用的网络，来讲述和验证本书介绍的许多网络性质。这些网络充分体现了网络科学所研究网络的多样性，包括社会网络、生物网络、技术网络和信息网络。线上课堂提供了本书所使用的这些数据集，读者可以使用它们作为学习网络科学的工具。

最后，对于使用英语之外的其他语言讲授本书的教授，在线课堂还会提供和持续更新本书多种语言的译本。

讲授网络科学

我曾经以两种不同的方式讲授过网络科学。第一次是一个完整学期的课程，选课的人主要是具有物理、计算机科学和工科背景的研究生及高年级本科生。第二次是一个为其3周的2学分的课，授课对象是具有经济学和社会科学背景的学生。这两次授课经验为本书的创作提供了重要借鉴：在那次完整学期的课程中，我讲授了书中的所有内

容，包括进阶阅读部分中的证明和推导；在那次较短的课程中，我只介绍了主要章节的内容，省略了进阶阅读部分以及和度相关性相关的章节。

在这两次授课中，同样重要的组成部分是下面将要介绍的课程作业和课程设计。

课后作业

在那次较长的课程中，我从每个章节后面列出的问题中选择一部分作为课后作业，测验学生对课程内容掌握的熟练程度以及解决问题的能力。整个课程中，我集中安排了两次课后作业。

维基作业

我们让每个学生选择一个与网络科学相关的概念或词语，创建一个维基百科页面对其进行介绍（图**10-1**　　）。这项作业的困难之处在于，学生们所选择的主题不仅不能是维基百科上已经存在的词条，还需要有足够的认可度，以便被维基接受。这项维基作业测试了学生以易于理解的百科全书式风格概述和提炼材料的能力，学生们还有可能通过完成这项作业成长为维基百科的贡献者。同时，这项作业丰富了维基百科上关于网络科学的内容，可以服务于整个网络科学领域。采用英语之外的语言讲授网络科学的教授，也可以考虑让学生用他们熟悉的语言撰写关于网络科学的维基页面。

维基作业

（1）选择一个和网络科学相关的关键词，确认该关键词尚未被维基百科收录。这里的“相关”很宽泛，可以选择一个技术概念（例如度分布）、一个网络相关的概念（例如恐怖分子网络）、一个网络的应用（例如金融领域的网络）、一位网络科学家，以及你认为和网络相关的任何事物。

（2）不要求原创。相反，你需要找出2 ~ 5个相关信息源（学术论文、书籍等），然后根据这些素材写出一个精炼的、自包含的百科全书式的概述，包括参考文献、图表、照片，尽可能全面。注意留意维基百科的版权声明等。

（3）把完成的页面上传到维基百科，然后把链接发给我们。维基百科不允许匿名添加新页面，因此你需要先在维基百科上创建一个账号。要确保你上传的页面不被维基百科的管理员删除。如果你上传的页面中概念不清晰、文献不完整或者风格不符合维基的要求，就有可能会被管理员删除。

（4）作业打分主要考查页面内容的可理解程度、客观度、自包含程度和准确程度。

图10-1　维基作业指南

社会网络分析

作为网络分析的热身，可以让学生们分析一下选课学生之间构成的社会网络。这需要事先做一些准备工作，还需要助教的协助。在第一节课上，教师可以收集一下选课学生的名单，让助教将名单打印出来发给每个学生，请学生从中标记出各自在选这门课之前就认识的人。为了让学生更方便地找出他们认识的人名，可以让每个学生做一个简单的自我介绍——这也有助于教师更好地了解选课学生。根据学生们标记后的列表，我们便可以构造出学生之间的社会网络，每个节

点还可以标记上学生性别和选课情况。在这门课上到一半的时候，将匿名化后的社会网络发给学生，让他们使用已经掌握的网络科学工具分析这个网络的性质。这给学生们提供了一个机会，去分析这样一个规模较小、亲身参与且容易理解的网络，从而有助于他们为完成最后的课程设计做好准备。这项作业可以使用软件来完成，以便学生们熟悉网络分析工具。

课程设计

课程设计是这门课最有趣的一部分，为学生们提供了一个综合利用所学知识来解决问题的机会。课程设计要求学生选择一个他们感兴趣的网络绘制出来，并对其进行分析。具体内容如下：

（1） 课程设计要求两人一组完成。条件允许的情况下，可以让学生们组成有差异的小组，譬如一个本科生和一个研究生结为一组，或者来自不同系的学生（例如生物系的学生和物理系的学生）结为一组。这样可以让学生们离开自己的舒适区，和与自己不同的人进行合作，而这正是交叉学科研究的常见情形。不用由教师来分组，而是鼓励学生们自己寻找合作伙伴。

（2） 课程开始几周之后，可以安排一次课，让学生们介绍一下课程设计的准备情况。每个组做一个不超过5分钟的报告，介绍一些他们选择的数据（图10-2）。我们要求学生自己收集数据，简单地下载一个已有的用于网络分析的数据是不能接受的。实际上，课程设计的目标之一就是让学生们体验一下选择网络时需要进行的折中。手工绘制网络是允许的。譬如，从烹饪书中找出食谱的配料绘制一个网络，或者绘制出某个小说或历史书中人物之间的关系。我们鼓励使用计算机工具绘制网络，譬如从万维网或数据库中收集数据，构造出一个网络。学生们需要对数据进行清洗，使其适合进行网络分析。例如，可以从维基百科中爬取数据，识别出作家、科学家或概念之间的关系。

准备一个不超过5分钟的报告：

- 介绍所选择的网络，对网络的节点和链接进行讨论。
- 介绍如何收集该网络的数据，并对网络大小进行估计，包括节点数 N 和链接数 L。要求节点数不少于 100。
- 介绍你选择该网络的理由。

图10-2　课程设计准备

（3）　需要强调的是，课程设计的目的是测试学生分析网络的能力。因此，学生们要重点关注数据中和网络相关的方面，避免被其他问题占用过多的精力而偏离目标。

（4）　课程结束时，要求学生们介绍各自的课程设计结果。根据选课人数的多少，可以安排一次或两次课来完成（图10-3）。

课程设计汇报

每个小组用 10 分钟时间汇报其课程设计完成情况。严格控制汇报时间。汇报的首页中需要写明学生的姓名、学生类别和来自哪个系。

介绍所选择的数据以及数据收集方法。选择一个数据条目进行介绍，使我们对数据有一个直观的认识。

测量：节点数 N 、链接数 L 以及二者随时间变化的情况（仅适用于动态网络），度分布，平均路径长度，集聚系数 $C(k)$，链接权重分布 $P(w)$（如果是带权网络）。对网络的社区结构进行可视化，探讨网络的鲁棒性、网络信息传播和度相关性等其他适合于你所选择网络的测量指标。

只是简单地进行测量还不够，你还需要对你得到的一些发现和理解进行阐述，可以尝试问以下问题：

- 你的期望是什么？
- 随机参照网络是什么？
- 如何将测量结果和你的期望进行对比？
- 你从这些测量的量中学到了什么？

评分依据：

- 网络工具的使用（完备性和正确性）；
- 使用网络工具从数据中进行信息抽取和发现的能力；
- 课程设计汇报的整体表现。

不需要提交文档，只需把报告通过邮件发给我们即可。

图10-3　课程设计汇报

在整个过程中，教师可能需要给学生多次反馈，包括维基关键词

的选择、组队搭档的选择、课程设计主题的选择等。为此，教师可以在每次课的最后留出10分钟时间询问学生并进行答疑：你们选好打算进行分析的网络了吗？这些网络的节点和链接分别是什么？你们知道如何获取这些数据吗？找到一起进行课程设计的搭档了吗？选择的维基关键词是什么？所选择的关键词是否已经被维基百科收录？收集了和你选择的关键词相关的文献了吗？这些问题的答案可能是“还没有”“不清楚”，或者是一些明确的答复。这样的询问和答疑，可以帮助学生评估他们各自的选择是否合适，帮助他们更清晰地理解自己的一些想法，帮助他们找到具有共同兴趣的课程设计搭档。一般情况下，几节课之后，每个学生都会找到自己的搭档，找到合适的课程设计主题并且选择了合适的维基关键词。

软件

我们可以安排一次课，让学生们尝试各种网络分析和可视化软件，例如Gephi、Cytoscape和NetworkX。如果课时比较多，还可以安排一次课，让学生们了解一些数值分析方法，包括数值拟合、对数分箱、网络可视化等。我们希望学生能把笔记本电脑带到课堂上，以便他们可以立即对这些工具和软件进行尝试。

电影之夜

在课外时间，我们会选择一个晚上作为“电影之夜”，观看安娜玛丽亚·塔拉斯的纪录片《连接》。这部时长一小时的纪录片介绍了网络科学的许多贡献者，讲述了网络科学这个领域无可辩驳的重要性。事先可以在整个校园里宣传“电影之夜”活动，让更多的人了解和参加。

客座讲者

在那次持续整个一学期的课程中，我们邀请网络科学领域的一些研究人员做了专题报告，讲授他们正在做的与网络科学相关的研究工作。这为学生们提供了一个能够了解该领域最前沿的研究的机会。这项活动通常会安排在课程的最后（也不尽然）——那时候大部分理论

工具都已经讲授完，学生的精力主要集中在课程设计上。专题报告同样是对外开放的，通常会为学生的课程设计带来启发和灵感。

为了辅助课程规划，图**10-4** 给出了我在一个完整学期开设网络科学课程时的课程表，那时候这本书尚未出版。

分数分布

（1）作业 1（家庭作业 1）：15%

（2）作业 2（家庭作业 2）：15%

（3）作业 3（课堂作业）：15%

（4）作业 4（维基作业）：15%

（5）课程设计准备：只有反馈，不占分数

（6）课程设计：40%

图**10-4** 课程考核体系

复杂网络：课程表

第一周

●第一次课：引言

●第二次课：第 1 章——图论

第二周

●第一次课：第 2 章——随机网络

●第二次课：第 2 章——随机网络

第三周

●第一次课：第 3 章——无标度性质

●第二次课：第 3 章——无标度性质

布置作业 1（引言至第 4 章的问题）

第四周

●第一次课：第 4 章——巴拉巴西 – 阿尔伯特模型

●第二次课：第 4 章——巴拉巴西 – 阿尔伯特模型

第五周

●第一次课：课程设计准备情况汇报

●第二次课：图表示、分箱方法、拟合方法

第六周

●第一次课：Gephi 和 Python；

收作业 1

布置作业 2：选课学生社会网络分析

●第二次课：特邀报告

第七周

●第一次课：第 5 章——演化网络

●第二次课：第 5 章——演化网络

第八周

●第一次课：特邀报告

收作业 2

●第二次课：第 6 章——度相关性

布置作业 3（第 5~9 章的问题）

第九周

第一次课：第 7 章——网络鲁棒性

布置作业 4：维基作业

第二次课：第 7 章——网络鲁棒性

第十周

第一次课：第 8 章——社区

第二次课：第 8 章——社区

电影之夜：安娜玛丽亚・塔拉斯的纪录片《连接》

第十一周

●第一次课：第 9 章——传播现象

●第二次课：第 9 章——传播现象

第十二周

●第一次课：特邀报告

●第二次课：第 9 章——传播现象

第十三周

●第一次课：特邀报告

●第二次课：公开课（课程设计讨论）

收作业 3

第十四周

●考试周：课程设计汇报（每组 10 分钟）

图10-5　课程大纲

4学分网络科学的课程大纲，每周设置了两次课。

10.2
致谢

任何一本书的撰写都是一次长途跋涉，《巴拉巴西网络科学》也不例外。撰写这本书几乎占用了我2011年至2015年之间的所有闲暇时间。而且，大部分内容的撰写工作都是我一个人独立完成的：在波士顿和布达佩斯的咖啡馆里，在全世界很多地方的凌晨，都能看到我努力写作的场景。尽管如此，这本书的完成也绝非我一个人的功劳：4年间，有很多人花费了大量时间和精力帮助我完成这一目标，也使我有机会和同事、朋友及实验室的伙伴一起讨论网络科学这一主题。另外，我通过互联网把所有章节的内容分享出来，收到了很多人的反馈和宝贵建议。在这一节，我想向在这次旅程的各个阶段给我提供过帮助的人表示感谢。



图10-6 数学团队

马顿·波什福伊负责本书的计算、模拟和测量工作。

公式、图和模拟

写一本教科书，需要所有事情都能够正常运转：按照书中内容能够推导出关键的公式，书中介绍的测量方法在真实数据上得到的结果应和理论预期吻合。要做到这一点，唯一的办法是对每次计算、测量和模拟进行重复和检查。这项伟大的工作大部分是由马顿·波什福伊完成的。马顿是从在我波士顿的实验室里做访问学生时开始这项工作的，随后他在匈牙利布达佩斯攻读博士学位期间，一直在继续着这项工作。他检查了所有的公式推导，进行了所有的模拟和测量，并准备了本书的图和表。有时候，准备一个图或表的工作量就相当于一个小的研究工作了。根据这些图和表的结果，我们弱化了对一些量的介

绍，将那些重要的量进行了重点介绍。马顿对网络科学文献的深入理解和认真细致的工作，帮助我们丰富了本书的内容。如果没有他孜孜不倦的投入，我可能无法保证本书的深度和可靠性。



图10-7 设计团队

莫罗·马蒂诺、加布里埃莱·穆塞拉和妮科尔·萨迈设计了每个章节的外观和图，为本书提供了优美一致的风格。

美工设计

莫罗·马蒂诺是我实验室里的一位数据可视化专家，他志在设计出一本符合大众审美的、具有视觉吸引力的书。他为每个章节设计了章首页，他设计的很多视觉元素贯穿全书。后来，莫罗去了IBM研究院，领导一个设计团队。加布里埃莱·穆塞拉接手了这本书的设计工作。他对本书的色彩进行了统一，设计了全书图表的基本元素，并且重新绘制了大部分的图。他和我们一起工作到2014年秋季，然后回到伦敦从事他梦想中的工作去了。本书的设计工作移交给了妮科尔·萨迈，她不辞辛苦地对整本书的设计进行了细致的润色。本书的网站是由金·阿尔布雷克特（Kim Albrecht）设计的，他目前在和莫罗一起设计用户在线阅读本书的体验。



图10-8 编辑团队

帕亚姆·帕斯内贾德、阿迈勒·侯赛尼和萨拉·莫里森负责这本书的编辑工作。

美工设计的一个重要部分是每个章节的章首页图片，这些图片向读者展现了网络和艺术的交相辉映。选择这些图片时，我从一些艺术

家和设计师那里获得了很多建议。感谢来自美国东北大学艺术设计系的伊莎贝尔·梅雷莱斯（Isabel Meirelles）和迪特马尔·奥芬哈勃（Dietmar Offenhuber）、来自哥伦比亚大学的马修·里奇（Mathew Ritchie）和来自旧金山艺术学院的梅雷迪思·特罗姆波（Meredith Tromble），感谢他们在我航行于艺术、数据和网络科学的边界时提供的帮助。



图**10-9**　校对和版权

菲利普·赫费尔是本书的最后一位编辑和第一位读者。布雷特·科蒙负责获取和管理本书中所有图的版权事宜。

文字工作：录入和编辑

我是一个传统的作家，在写作时不使用计算机，而是用铅笔来完成。因此，我需要有人来协助我把写出的东西录入计算机。最初，这项工作是由萨布丽娜·拉贝洛（Sabrina Rabello）和盖伦·威尔克森（Galen Wilkerson）负责的。后来，这项繁重的工作落到了另外三个人的肩上。第一年，帕亚姆·帕斯内贾德（Payam Parsinejad）在协助我进行写作。在他把主要精力转移到研究工作上之后，曾经选修过我的网络科学课程的一名学生阿迈勒·侯赛尼加入进来，并且一直坚持到了最后。另外，我之前在波士顿的助手萨拉·莫里森在搬迁到意大利卢卡之后，也加入了这项工作。她及时准确的录入和编辑对于本书的写作非常重要。

在本书内容发布到网站上之前，每个章节都由菲利普·赫费尔做了

一遍最后检查。菲利普是在访问我的实验室时加入这项工作中来的，在回到柏林创建自己的实验室之后仍然负责这项工作。菲利普对每个章节内容的各个方面进行检查，为这本书把好最后一关，他也是本书的第一位读者。

布雷特·科蒙（Brett Common）负责确保书中使用的所有图片都得到了许可。这是一项非常重要的工作，其工作量和工作难度也都非常大。



图**10-10** 家庭作业

罗伯塔·西纳特拉为本书的每个章节设计了家庭作业。

家庭作业

每个章节最后的家庭作业是由罗伯塔·西纳特拉设计的。作为我实验室里的一名研究人员，罗伯塔和我一起讲授了2014年秋季的网络科学课程，在教学过程中，她帮助我指出并修正了课件中的许多拼写错误。

学界帮助

在撰写这本书的过程中，有很多科学家和学生给出了有益的评论和建议，进行了推荐和解读，并提供了关键素材。虽然难以面面俱到，但我会尽量列出对本书提供过帮助的全部人员。

宋朝鸣（Chaoming Song）帮助我估计无标度网络的度指数，并协助我收集级联失效相关的文献。数学家安德烈·乔考（Endre Csóka）帮助我清楚地阐述了波罗巴斯模型的细节。我还从一些讨论中受益匪浅：和雷萨·德索萨（Raissa D’Souza）讨论了优化模型，和吉内斯特拉·比安科尼（Ginestra Bianconi）讨论了适应度模型，和伊

丽莎白·劳沃斯·里根（Erzsébet Ravasz Reagan）讨论了劳沃斯算法。亚历山德罗·维斯皮那尼在扩散过程和度相关性方面提供了大量帮助。马里安·博古纳（Marian Boguña）绘制了空手道俱乐部网络。沈华伟（Huawei Shen）对论文的未来引用次数进行了计算。盖尔盖伊·帕拉和陶马什·维切克帮助我理解和使用CFinder软件，马丁·罗斯瓦尔（Martin Rosvall）提供了InfoMap算法的一些关键资料。盖尔盖伊·帕拉、苏内·莱曼（Sune Lehmann）和圣·福尔图纳托（Santo Fortunato）为社区识别那一章提供了重要建议。安庸烈（Yong-Yeol Ahn）帮助我写出了扩散现象的第一版。拉米斯·莫瓦萨（Ramis Movassagh）、佐山世纪（Hiroki Sayama）和锡德·雷德纳（Sid Redner）对本书的一些章节提供了细致的反馈，凯特·克罗格斯（Kate Coronges）帮助我对前4章进行了润色。

出版事宜

西蒙·卡珀兰（Simon Capelin）——我在剑桥大学出版社长期合作的编辑，一直鼓励我写这样一本书。这本书在完成之前经过了多次推迟，但他一直都很耐心。在出版过程中，罗欣·芒内利（Róisín Munnelly）也提供了大量帮助。

单位致谢

最后，我要感谢几家单位。如果没有它们提供的环境和支持，这本书可能无法完成。首先，我要感谢的是我在东北大学的一些领导，包括校长约瑟夫·奥恩（Joseph Aoun）、教务长史蒂夫·迪雷克托（Steve Director）、院长默里·吉布森（Murray Gibson）和拉里·芬克尔施泰因（Larry Finkelstein），以及系主任保罗·钱皮恩（Paul Champion），他们是网络科学的倡导者，使网络科学在东北大学成为一门跨学科的主修课程。在他们的大力支持下，学校聘用了许多研究网络的著名科学家。这些科学家来自很多不同的领域——包括物理、数学、社会学、政治学、计算机科学和健康科学，使东北大学成为网络科学领域的领跑者。在他们的倡导和支持下，东北大学设立了

网络科学博士学位计划，创建了由维斯皮那尼领导的网络科学研究所。

我在哈佛医学院布里格姆女子医院的网络医药部和丹娜 - 法伯癌症研究所的系统生物学癌症中心的职位，为将网络科学应用到细胞生物学和医学领域提供了一个窗口。多谢丹娜 - 法伯癌症研究所的马克·维达尔（Marc Vidal）和来自布里格姆女子医院的乔·洛斯卡尔佐（Joe Loscalzo），他们在该领域的实践为本书提供了很多支持。

我在中欧大学的客座职位以及在那里开设的网络科学夏季课程，使我接触到很多有着经济学和社会科学背景的学生。这段经历奠定了这本书的雏形。富有远见的巴拉兹·韦德赖什（Balázs Vedres）将网络科学带到了中欧大学，乔治·索罗斯（George Soros）说服我加入中欧大学。校长约翰·沙特克（John Shattuck）、教务长法卡斯·卡塔林（Farkas Katalin）和利维乌·马泰（Liviu Matei）的大力支持为中欧大学开设网络科学课程、设立网络科学博士学位的计划铺平了道路。

最后，我要感谢网络科学开始的地方。当我还是一个助理教授时，圣母大学为我提供了支持和能够让我尝试不同事物的环境。非常感谢我的助理苏珊娜·奥莱沃（Suzanne Aleva），从圣母大学到东北大学，她跟随我的实验室不辞劳苦地工作了十余年，为我营造了很好的科研环境，使我可以不受打扰地专注于科学本身。

未来，属于终身学习者

我这辈子遇到的聪明人（来自各行各业的聪明人）没有不每天阅读的——没有，一个都没有。巴菲特读书之多，我读书之多，可能会让你感到吃惊。孩子们都笑话我。他们觉得我是一本长了两条腿的书。

——查理·芒格

互联网改变了信息连接的方式；指数型技术在迅速颠覆着现有的商业世界；人工智能已经开始抢占人类的工作岗位……

未来，到底需要什么样的人才？

改变命运唯一的策略是你要变成终身学习者。未来世界将不再需要单一的技能型人才，而是需要具备完善的知识结构、极强逻辑思考力和高感知力的复合型人才。优秀的人往往通过阅读建立足够强大的抽象思维能力，获得异于众人的思考和整合能力。未来，将属于终身学习者！而阅读必定和终身学习形影不离。

很多人读书，追求的是干货，寻求的是立刻行之有效的解决方案。其实这是一种留在舒适区的阅读方法。在这个充满不确定性的年代，答案不会简单地出现在书里，因为生活根本就没有标准确切的答案，你也不能期望过去的经验能解决未来的问题。

湛庐阅读App：与最聪明的人共同进化

有人常常把成本支出的焦点放在书价上，把读完一本书当作阅读的终结。其实不然。

时间是读者付出的最大阅读成本
怎么读是读者面临的最大阅读障碍
“读书破万卷”不仅仅在“万”，更重要的是在“破”！

现在，我们构建了全新的“湛庐阅读”App。它将成为你“破万卷”的新居所。在这里：

- 不用考虑读什么，你可以便捷找到纸书、有声书和各种声音产品；
- 你可以学会怎么读，你将发现集泛读、通读、精读于一体的阅读解决方案；
- 你会与作者、译者、专家、推荐人和阅读教练相遇，他们是优质思想的发源地；
- 你会与优秀的读者和终身学习者为伍，他们对阅读和学习有着持久的热情和源源不绝的内驱力。

从单一到复合，从知道到精通，从理解到创造，湛庐希望建立一个“与最聪明的人共同进化”的社区，成为人类先进思想交汇的聚集地，与你共同迎接未来。

与此同时，我们希望能够重新定义你的学习场景，让你随时随地收获有内容、有价值的思想，通过阅读实现终身学习。这是我们的使命和价值。

湛庐CHEERS

湛庐阅读App玩转指南

湛庐阅读App结构图：



三步玩转湛庐阅读App：

读一读 ▼

湛庐纸书一站买，
全年好书打包订

听一听 ▼

泛读、通读、精读，
选取适合你的阅读方式

扫一扫 ▼

买书、听书、讲书、
拆书服务，一键获取



App获取方式：
安卓用户前往各大应用市场、苹果用户前往 App Store
直接下载“湛庐阅读”App，与最聪明的人共同进化！

湛庐CHEERS

使用App扫一扫功能，遇见书里书外更大的世界！



快速了解本书内容，
湛庐千册图书一键购买！

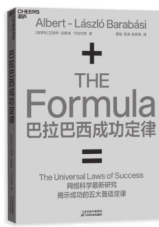
大咖优质课、
献声朗读全本一键了解，
为你读书、讲书、拆书！

你想知道的彩蛋
和本书更多知识、资讯，
尽在延伸阅读！

延伸阅读

《巴拉巴西成功定律》

- 全球复杂网络研究权威、“无标度网络”奠基人艾伯特 - 拉斯洛·巴拉巴西首度用故事揭示人类成功5大科学定律！宜警觉！可实践！能复制！
- 全覆盖、无死角的数据支持，汇集100多位世界顶尖的科学家、社会学家、心理学家、商业界领袖的最新研究成果，在世界范围内首次揭示了成功背后那张“看不见的网”。



使用“湛庐阅读”App，
“扫一扫”获取本书更多精彩内容。

ISBN 978-7-5576-7280-5

《链接》

- 艾伯特 - 拉斯洛·巴拉巴西的成名之作，同时也是复杂网络的奠基之作、社交网络的入门之作。作者巴拉巴西第一个证明了，我们不是生活在随机世界里，真实网络是无标度的。
- 《链接》一书可以被视为复杂网络的基石、大数据时代的开端。书中追溯了网络的数学起源，分析了社会学家在此基础上得出的研究成果，为我们的网络研究提供了一个全新的视角。



《爆发》

◎ 一本超越《黑天鹅》的惊世之作。如果说塔勒布认为人类行为是随机的，都是小概率事件，是不可以预测的；那么全球复杂网络权威巴拉巴西则认为，93%的人类行为是可以预测的。

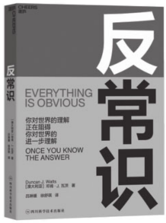
◎ 《爆发》一书揭开了人类行为背后隐藏的模式“爆发”，提出人类日常行为模式不是随机的，而是具有“爆发性”的。



《反常识》

◎ 小世界网络之父、微软研究院首席科学家邓肯·瓦茨的颠覆式新作！本书以互联网为“望远镜”探索社会运行的规律，明确界定了哪些事情适合用常识来判断和解决，哪些事情适合用反常识来处理。

◎ 邓肯·瓦茨根据最新的网络科学研究以及商业、政治和日常生活中比比皆是、内容翔实的决策案例，揭晓了常识带来的4大误区和反常识带来的3大红利，并教授我们学会用反常识思维认识并应对这个复杂的世界，跳出常识陷阱，优化工作与生活。



使用“湛庐阅读”App，
“扫一扫”获取本书更多精彩内容。

ISBN 978-7-5364-9541-8